W203 Lab 1: Candidate Debt Exploratory Data Analysis

Eric Hulburd, Mitch Stepleton, Yulia Zamriy

January 30th, 2018

1. Introduction

As members of the Washington State Election Commission, it is our responsibility to ensure that all elections that occur within the State are free, fair, and transparent. It seems that now, more than ever, a clear understanding of how campagins are financed is necessary to meet those responsibilities. To that end, we have been asked to perform an exploratory analysis of candidate debt filing data from the 2012 election cycle. The goal of this exploratory analysis is to better understand how campaign characteristics are related to the financial debts of candidates.

The analysis is broken into the following 5 sections, with findings and recommendations contained in the conclusion:

1. **Introduction**
2. **Univariate Analysis**
3. **Analysis of Key Relationships**
4. **Analysis of Secondary Effects**
5. **Conclusion**

1.a. Loading the Data

```
CandidateDebt <- read.csv("CandidateDebt.csv",
                          stringsAsFactors = FALSE)
#str(CandidateDebt)
table(CandidateDebt$amount)
```

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

Immediately upon loading the data, it became clear that the columns in the original file had been mislabelled, with each column header shifted one space to the left starting with legislative district. The final column at the end of the file, vendorzip, did not have values in the file. To address this, we pulled the header from the file, corrected the column locations, and saved the result to the vector var_names_corrected.

```
# 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)])
```

We reloaded the data and we replaced the header with the corrected values so that columns were correctly labeled.

```
# 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"))
```

1.b. Describing the Data

The relabeled dataset consisted of 1,043 rows, and 28 columns. Though all columns were initially formatted as character strings, conceptually the data contains:

- 24 strings Fields:
  - **id**: An iternal identifier that corresponds to a single expenditure record.
  - **reportnumber**: An identifier used for tracking the individual form.

  - **origin**: This field shows from which filed report-type the data originates.
  - **filerid**: The unique id assigned to a candidate.
  - **filertype**: ndicates if this record is for a candidate.
  - **filername**: he candidate or committee name as reported on the candidates registration.
  - **firstname**: his field represents the first name, as reported by the filer.
  - **middleinitial**: his field represents the middle initial, as reported by the filer.
  - **lastname**: his field represents the last name, as reported by the filer.
  - **office**: he office sought by the candidate.
  - **legislativedistrict**: The Washington State legislative district.
  - **position**: The position associated with an office.
  - **party**: he political party as declared by the candidate on their registration.
  - **jurisdiction**: The political jurisdiction associated with the office of a candidate.
  - **jurisdictioncounty**: The county associated with the jurisdiction of a candidate.
  - **jurisdictiontype**: The type of jurisdiction this office is: Statewide, Local, etc.
  - **recordtype**: This field designates the item as a debt.
  - **code**: The type of debt.
  - **description**: The reported description of the transaction.
  - **vendorname**: The name of the vendor or recipient's name.
  - **vendoraddress**: The street address of the vendor or recipient.
  - **vendorcity**: The city of the vendor or recipient.
  - **vendorstate**: The state of the vendor or recipient.
  - **vendorzip**: The zip code of the vendor or recipient.
    - *Note: No values in dataset original dataset.*
- 3 Date Fields:
  - **fromdate**: The start date of the period for the report on which this debt record was reported.
  - **thrudate**: The end date of the period for the report on which this debt record was reported.
  - **debtdate**: The date that the debt was incurred.
- 2 Numeric Fields:
  - **electionyear**: The election year in the case of candidates.
  - **amount**: The amount of the debt incurred or order placed.

```
dim(CandidateDebt)
```

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

```
#summary(CandidateDebt) - commented out for readability.  Detailed summary information available upon request.
```

1.c. Evaluating and Addressing Data Quality

1.c.i. Correcting Data Types

The first step of our data quality assessment was casting variables to their correct data type in R.

Casting date fields from character strings to dates:

```
# 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")
```

Casting amount from character string to numeric:

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

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

1.c.ii. Addressing #N/A Values

Preliminary analysis revealed 56 rows for which the amount variable as '#N/A':

```
# 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
```

Closer inspection of these 56 rows revealed that they were missing data in all columns except filer name and office they run for. Additionally, all 56 values related to a single candidate in the dataset which suggests a data processing issue associated with those entries. Given that these rows did not contain values for the debt amount and did not contain a unique value of office not otherwise seen in the data, we excluded them from our analysis.

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

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

```
# 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
```

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

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

```
# 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
```

1.c.iii. Removing Unnecessary Columns

With basic scrubbing and and recasting of variables completed, our next step was to compose the dataset that would be used for the analysis. Several of the variables within the dataset contained information that were either redundant or unnecessary and as such were removed to make the dataset easier to work with.

Excluded variables:

- **origin**: Only one value ("B.3") for all records.
- **filertype** Only one value ("Candidate") for all records.
- **filername** & **firstname** & **middleinitial** & **lastname**: Analysis will use filerid as a unique candidate identifier.
- **position** and **position2** (values are not clear and were inconsistent in raw format)
- **electionyear**: Only one value ("2012") for all records.
- **recordtype**: Only one value ("DEBT") for all records.

```
# 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)
```

After removing redunant columns, the final dataset to be used in data exploration and analysis contained 987 rows and 19 columns. The below output serves as a check that all variables of interest are in the correct format and there are no apparent missing values.

```
dim(CandidateDebtSub)
```

```
## [1] 987  19
```

```
#summary(CandidateDebtSub) -- commented out for readibility.  More Detailed information is available upon reques
```

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

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

1.c.iiii. Testing Variable Consistency

Using the trimmed dataset, we performed additional quality checks on variables of interest looking for anything notable or out of the ordinary. Specifically, for each variable, we counted the distinct values it had in the dataset per candidate. Given that this data is only for a single election cycle, we would expect values such as **party**, **office**, and **jurisdiction** to only have a single value per candidate.

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

#sapply(aggr_comb[, -1], table)
summary(aggr_comb[, -1])
```

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

Unfortunately, this analysis revealed that several fields otherwise of interest to us contain values that are most likely inaccurate.

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. Unsure of whether these instances represented an inaccuracy in data reporting and collection, or whether many of these candidates were running simultaneous campaigns, we decided to simplify the data set and focus solely on fields that did not suggest the possibility of inaccurate candidate identification or multiple campaigns within the same season. We, therefore, will exclude these variables specified below and provide guidance for improving data capture and quality moving forward.

2. Univariate Analysis

We conducted univariate analyses on the variables we believed we could reliably explore as discussed above. The objective of this subset of analysis is to better understand the distribution of each variable and to identify specific variables that may be informative in a bivariate analysis.

2.a Amount of Debt Incurred

The amount variable in our dataset is the transacted amount or amount of debt incurred by the candidate on a particular occassion. This dataset can give us insight into the distribution of the quantities candidates report to spend. For this analysis, we use the pre-processed `amount_num` field which has properly been parsed for numeric analysis.

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

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

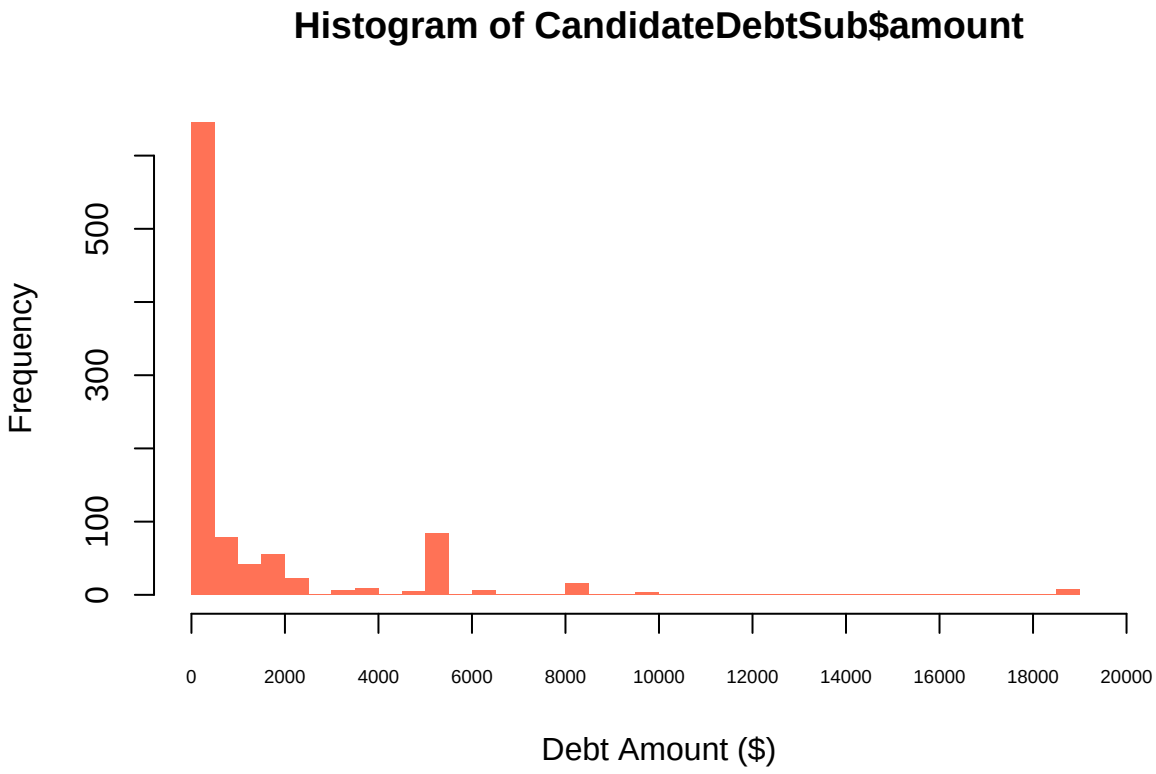
```
sd(CandidateDebtSub$amount_num)
```

```
## [1] 2494.271
```

As evidenced above, there is a broad range of transactions reported by candidates - from $3.24-$19,000. While $19,000 is a large sum, it is assuringly more on the order of a hefty consulting fee rather than a new Porsche or vacation home in the Bahamas.

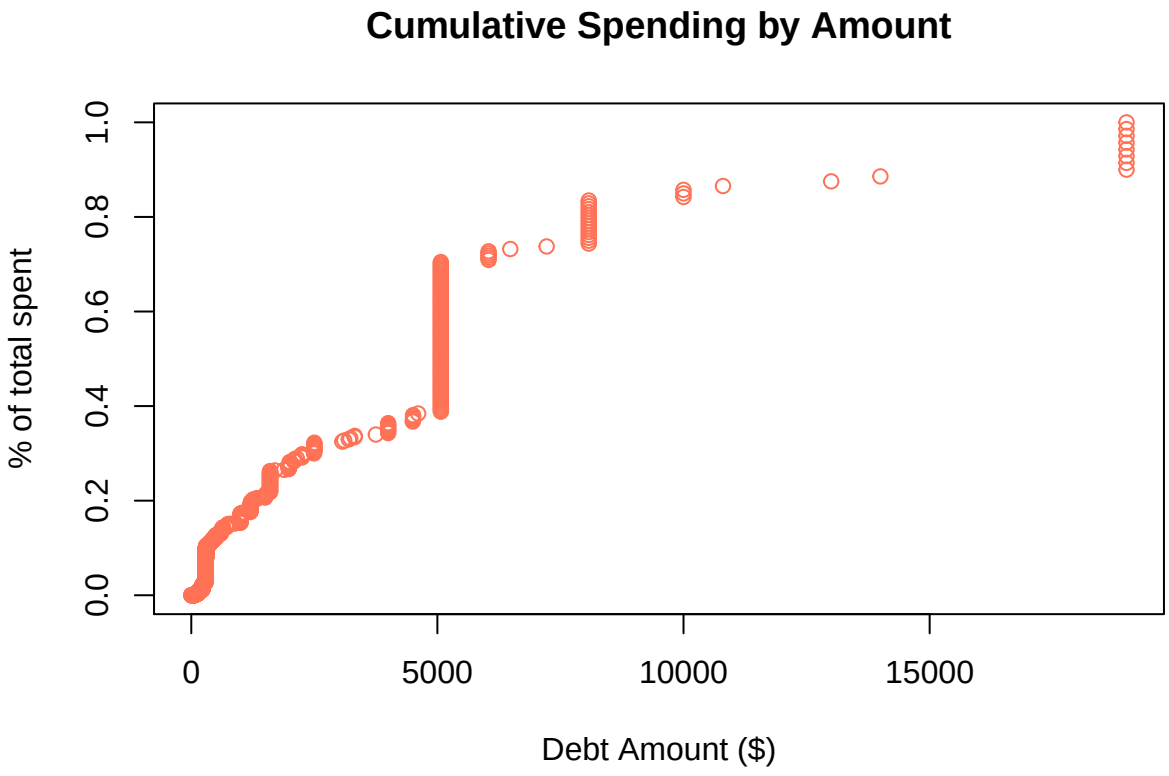
Interestingly, the standard deviation of the debt incurred, $2494.27, is larger than the mean, $1347.43, while the third quartile, $1210.50, is less than the mean. This suggests a highly skewed distribution and, indeed, when we look at the histogram, the distribution is skewed far to the left.

```
# 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)
```



Despite the skew towards reported number of smaller transactions, it is interesting to view this in relation to total amount spent by candidates.

```
total_spent = sum(CandidateDebtSub$amount_num)
plot(sort(CandidateDebtSub$amount_num), cumsum(sort(CandidateDebtSub$amount_num)) / total_spent, xlab="Debt Amou
```



This chart above goes to show how the frequency of the distribution amount may often be deceptive in terms of where most money is being spent by candidates and helps us hone in on the following.

```
# Share of debt with less than $500 per report
sum(subset(CandidateDebtSub, amount_num < 500)$amount_num) / total_spent
```

```
## [1] 0.1250873
```

```
# Share of debt with more than $5000 per report
sum(subset(CandidateDebtSub, amount_num >= 5000)$amount_num) / total_spent
```

```
## [1] 0.6156793
```

```
# Number of reports with $5070.14 in debt
sum(CandidateDebtSub$amount_num == 5070.14)
```

```
## [1] 84
```

```
# Share of debt with more than $19000 per report
sum(subset(CandidateDebtSub, amount_num >= 19000)$amount_num) / total_spent
```

```
## [1] 0.1142936
```

```
# Number of reports with $19000 in debt
sum(CandidateDebtSub$amount_num == 19000)
```

```
## [1] 8
```

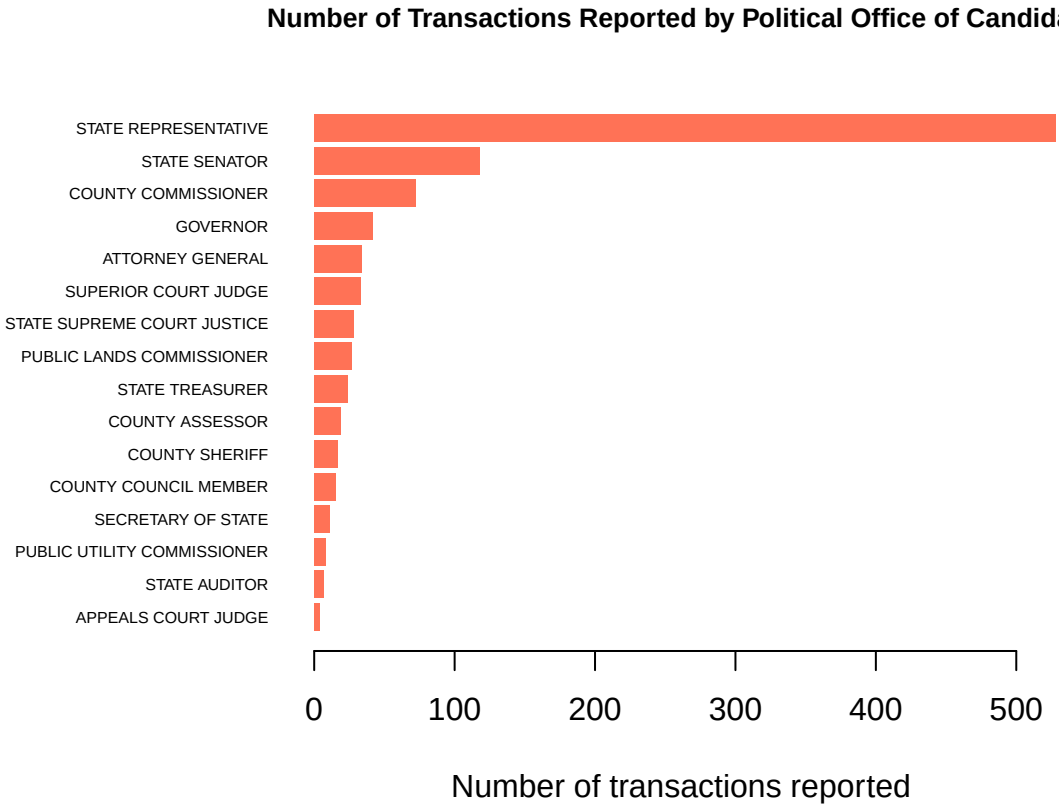
While, just about 2/3 of the reported transaction amounts were under $500, those transactions only amounted to about 12% of the total spent.

Conversely, those transactions over $5,000 (over a full standard deviation above the mean of the distrubtion) and $19,000 accounted for 62% and 11%, respectively, of all money spent.

Curiously, we also see that eight of these reported amounts were for $19,000 exactly and 84 were for $5,070.14. We will consider if these amounts were related to a particular vendor-service or if there were financial disincentives to breach these marks.

2.b Office Sought by Candidate

```
par(mar=c(4,8,4,5))
barplot(sort(table(CandidateDebtSub$office)),
            horiz=TRUE,
            las = 1,
            cex.names = 0.5,
            xlab='Number of transactions reported',
            col='coral1',
            border=NA)
title('Number of Transactions Reported by Political Office of Candidate',
      cex.main = 0.8)
```



As one would suspect the highest number of transactions were reported by candidates for state representative and senator. This, of course, makes sense since state Congress would have the highest number of positions and, therefore, candidacies That candidates for the governorship reported the third highest number of transactions, suggesting (a) it is a popular office to which many people aspire to or (b) candidates spend more frequently on their candidacy. We will need to look at the office relationship with the filerid and transaction amount.

2.c Vendor City and State

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

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

```
prop.table(table(CandidateDebtSub$vendorstate))
```

```
##
##                      CA          DC          TX          WA
## 0.025329281 0.010131712 0.101317123 0.005065856 0.858156028
```

It comes as no surprise that the over 85% of the reported transactions occurred within Washington state where the candidates must campaign and appeal to voters. It should also be no surprise to see that Washington DC, where national political parties are headquartered and lobbyists are abundantly present, is the second most frequent state of transaction. Some candidates may additionally have ventured out to larger states with political party headquarters for support for their campaign. Looking at the relationships of spending in these states, as well as the office run for, may give some indication as to why some candidates spend more outside the states borders than others.

2.d Transaction Type and Description

```
# 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)
```

The analysis of the transaction type was very straightforward. Most of the time, candidates neglected to classify the type of the transaction. The answer to the question "why?" is beyond the scope of this exploratory analysis, however, examining the relationship of the transaction type with the amount may indicate whether entering that information seemed too tedious for small amounts of money or whether candidates were hiding the type of their larger transactions.

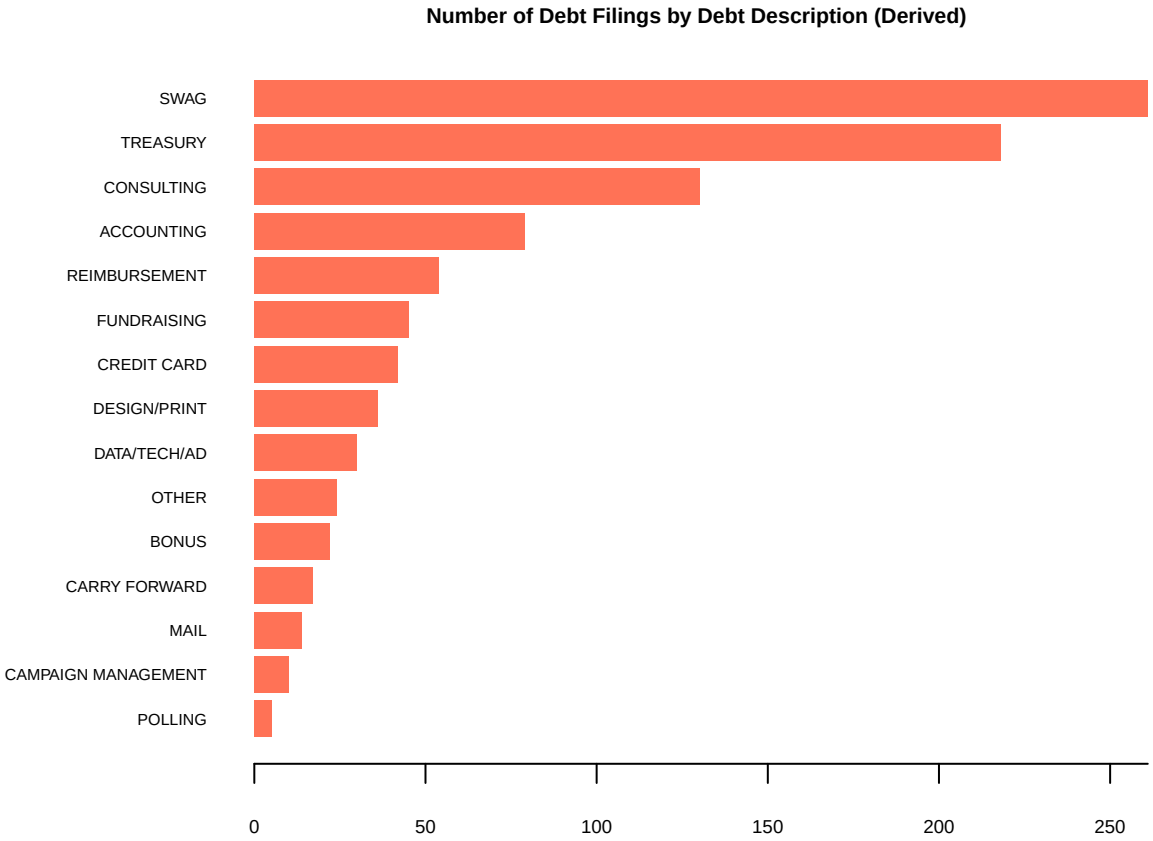
While the code field required little pre-processing, its value was limited due to the number of un-filled rows. Conversely, the description field does not have blanks, but is at a level of granularity that makes it unwieldly for this analysis. To solve these issues, we created a derived variable that categorizes candidate debt based on the filing description. We grouped the descriptions into the following categories: treasury, campaign management, fundraising, carry forward, reimbursement, accounting, bonus, design and print, polling, credit card, consulting, swag, mail, and date, technology, and ads.

Note that these are in no way comprehensive nor mutually exclusive. We grouped transactions into these categories by searching for keywords within the transaction description. We included this rudimentary analysis in order to see if there were any obvious trends. Any sort of inferential analysis would require more thorough research and perhaps more sophisticated text searching.

```
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"
```

```r
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"))
```

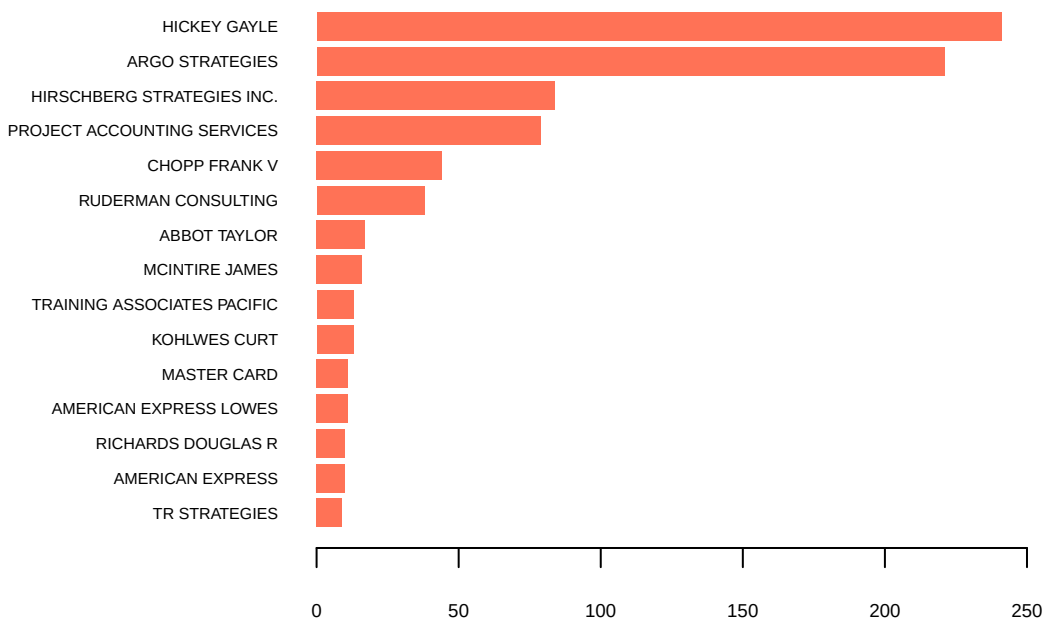
```r
aggr_descr <- aggregate(amount_num ~ description_aggr, data = CandidateDebtSub, sum)
aggr_descr <- aggr_descr[order(aggr_descr$amount_num),]

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)
```



While it appears that there was a high volume of transactions that we categorized as swag (ie campaign materials), there were also a high number of treasury and consulting transactions, which we would expect. We will explore this relation to transaction amount in the upcoming section.

2.e Vendor Name

```
par(mar=c(4,10,4,4))
vendorTable <- sort(table(CandidateDebtSub$vendorname),decreasing = TRUE)
barplot(sort(vendorTable[1:15]),
        horiz = TRUE,
        las = 2,
        cex.names=0.5,
        cex.axis = 0.8,
        xlim = range(0:250),
        axes = FALSE,
        col='coral1',
        border=NA)
axis(1, at = seq(0, 250, 50),
     cex.axis = 0.6,
     las = 1)
title('Number of Debt Filings by Vendor',
      cex.main = 0.9)
```

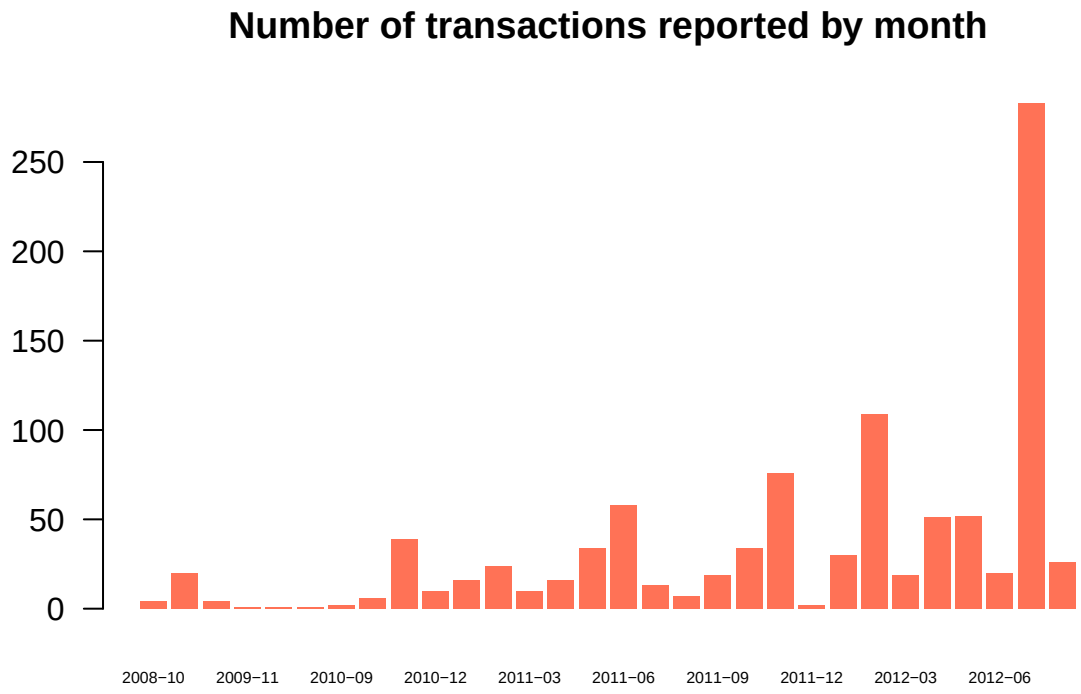


Analysis of the vendor name is very straightforward. We simply took a look at the most frequent vendor names so we can use these for further analysis to identify where candidates are spending most of their money and whether these vendors are integrated across a broad spectrum of political services and goods or if they are specialized.

2.f Transaction and Report Dates

As mentioned in the introduction, we parsed all of the date fields to R date objects. The data extends from October 2008 until August 2012. Because political campaign cycles are, fortunately, not annual, it is unlikely we would see the same spending patterns year after year. We can run a rudimentary analysis with a histgram to see when most the most frequent months and years were for campaign transactions.

```
barplot(table(sort(format(CandidateDebtSub$debtdate, format="%Y-%m"))),
            las = 1,
            cex.names = 0.5,
            col='coral1',
            border=NA,
        main = 'Number of transactions reported by month')
```



These charts show us the following: 1. Reported spending frequency peaked in July 2012 and ended August 2012. 2. There were no clear trends for any month year-to-year. 3. Overall, spending increased year-to-year from 2008 to 2012.

3. Analysis of Key Relationships

3.a Number of Candidates and Average Debt per Candidate By Office

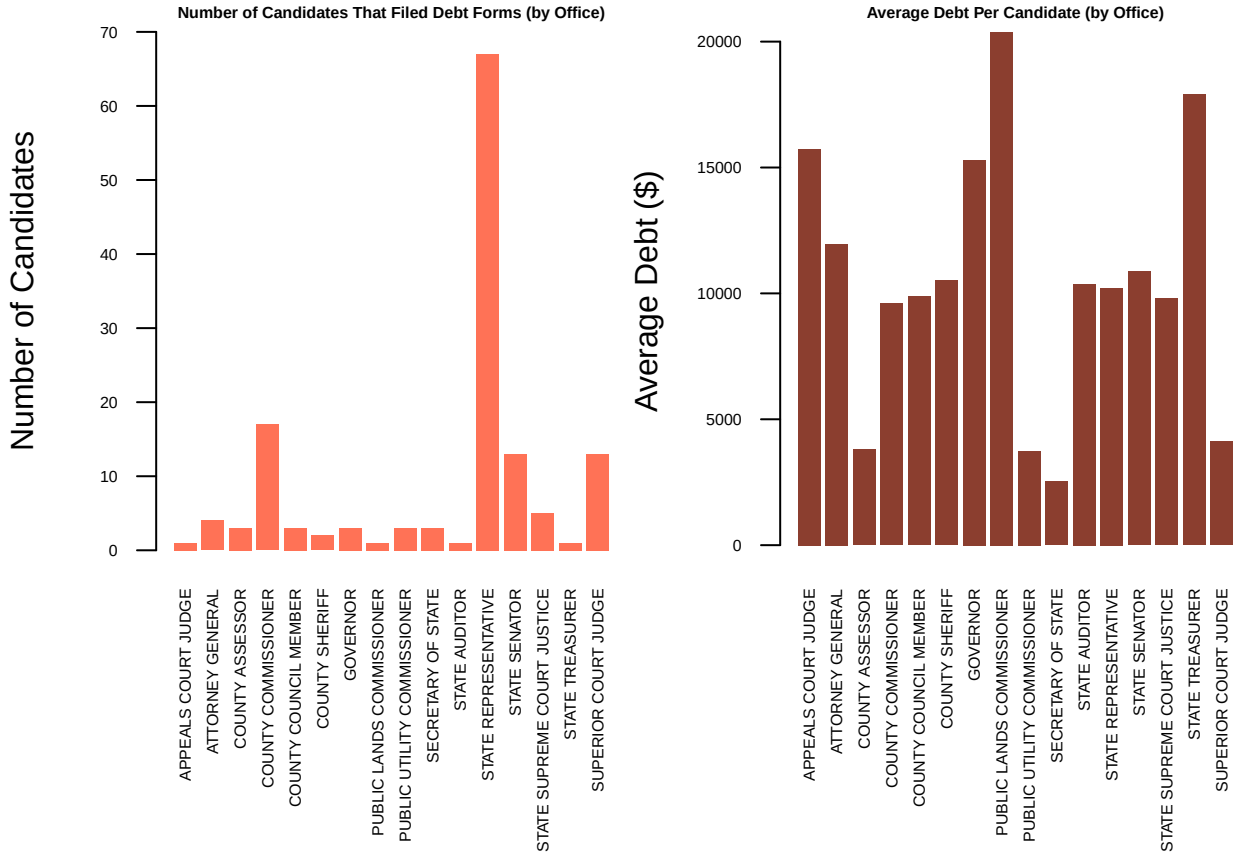
As we covered in the earlier sections, the final analytical dataset contains data on 140 candidates that filed debt reports. Based on the first chart in the figure below, almost half of the candidates were running for "State Representative". "County Commissioner", "State Senator" and "Superior Court Judge" offices had between 15 and 20 candidates, while the rest of the offices had less than five. Is it possible that not all candidates accumulated debt during the eleciton and, hence, did not have a record in this report? That would be an important piece of information to add to make this analysis more comprehensive.

Another piece of data required for a complete picture is the number of seats available at each office. We would use it to normalize the metric and understand if there was a disproportionate number of candidates filing debt reports in some offices vs. the others.

The second chart in the figure below displays average debt per candidates by Office during the election. It is interesting to note that the offices with the highest average debt per candidate (Public Lands Commissioner, State Treasurer and Appeals Court Judge) had the smallest number of candidates in the election (assuming they all filed debt reports). However, we cannot make any conclusions based only on these few data points. Is it just a coincidence? Are the campaigns for these offices more expensive, excluding other potential candidates with financial resources?

```
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(8,4,1,0),
    oma = c(0,0,0,0),
    mfrow = c(1,2))
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 = 0.5)

par(mar = c(8,4,1,0))
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 = "coral4")
title("Average Debt Per Candidate (by Office)",
      cex.main = 0.5)
```




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

3.b Timing Analysis of the Debt

Let's take a look on how debt occured in the months leading to the election (across all candidates and all offices). Overall, the first debt report available in the file was dated 46 months prior to the election. However, debt amounts were relatively small until 7 months prior to the election (with two exceptions of 18 and 14 months). Moreover, the largest amount of debt was incurred 6 months prior to the election. We would need to investigate why that was. Were there certain events in the campaign calendar that required significant spend at that time? Or was it a consequence of individual candidates spending patterns? One of the ways to approach this question would be to compare this patter across multiple campaigns.

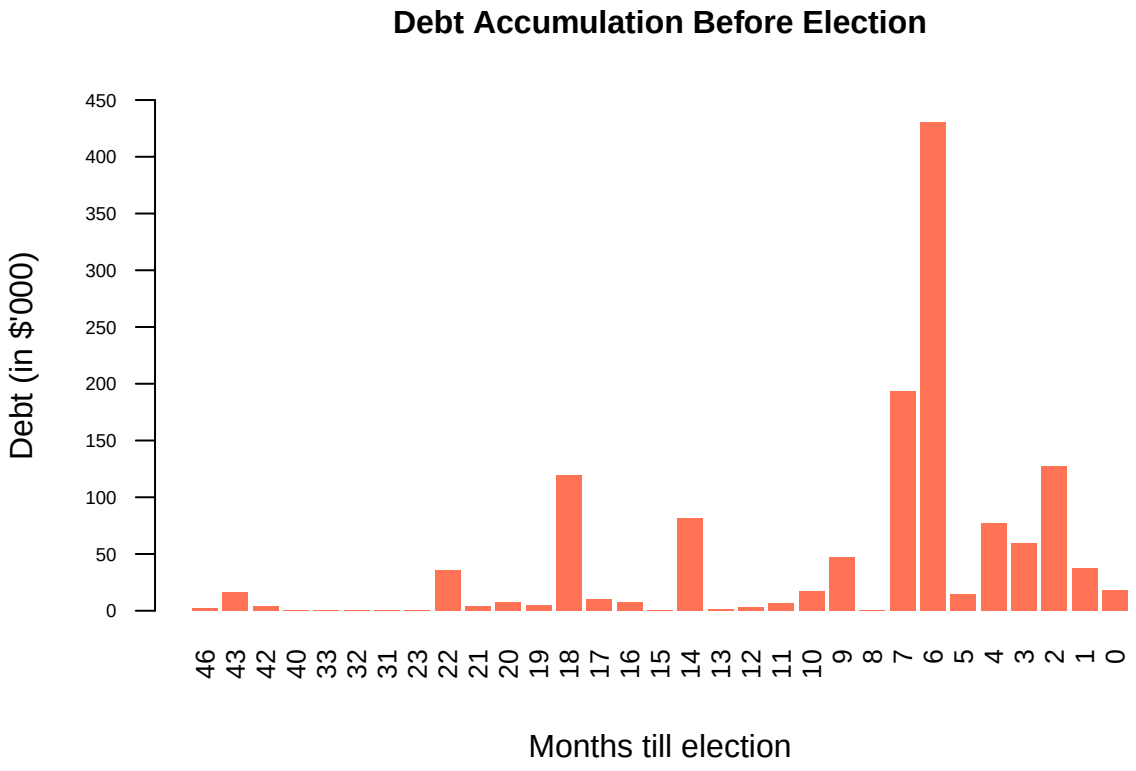
```r
# Number of months before election the debt occured
CandidateDebtSub$weeksindebt <-
  round(difftime(max(CandidateDebtSub$debtdate), CandidateDebtSub$debtdate, units = "weeks"))
CandidateDebtSub$monthsindebt <-
  round(CandidateDebtSub$weeksindebt / 52 * 12)
CandidateDebtSub$monthsindebt <-
  as.numeric(CandidateDebtSub$monthsindebt)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

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

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

```r
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)
```

This provides a different picture of the timing of the spending. While the univariate analysis suggested that the number of transactions picked up in July 2012, this analysis shows that, in fact, the total transaction volume actually peaked six months before the election.

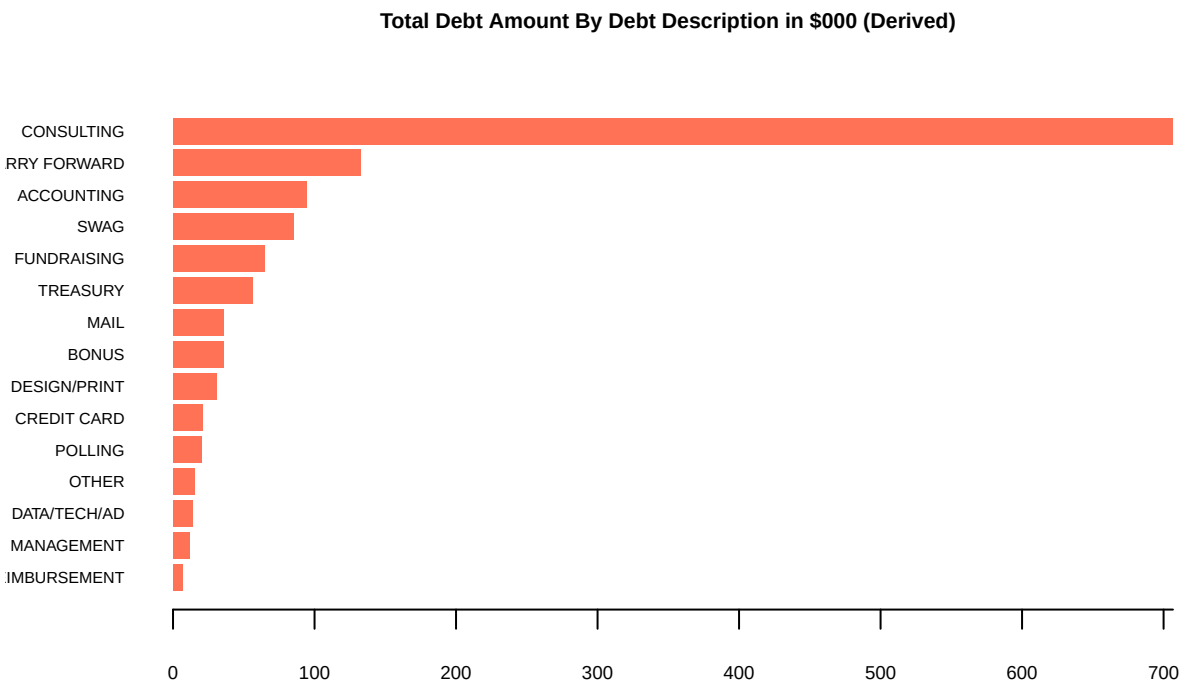
3.c Analysis of Debt by Type

One of the key questions we tried to investigate was how candidates spent campaing money. This was captured in the description variable. We aggregated it into fewer groups for analysis purposes.

We can see from the first chart below, that most of the reports were related to campaign materials. And while consulting related debt reports were only #3 by their abosolute quantity, this category was by far the largest in terms of amount spent. Is it common practice to spend this much on consulting? We would need to compare a few campaigns to make any conclusions.

```
# description_aggr

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

#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, 800000, 100000),
     cex.axis = 0.6,
     labels = seq(0, 800, 100),
     las = 1)
title('Total Debt Amount By Debt Description in $000 (Derived)',
```

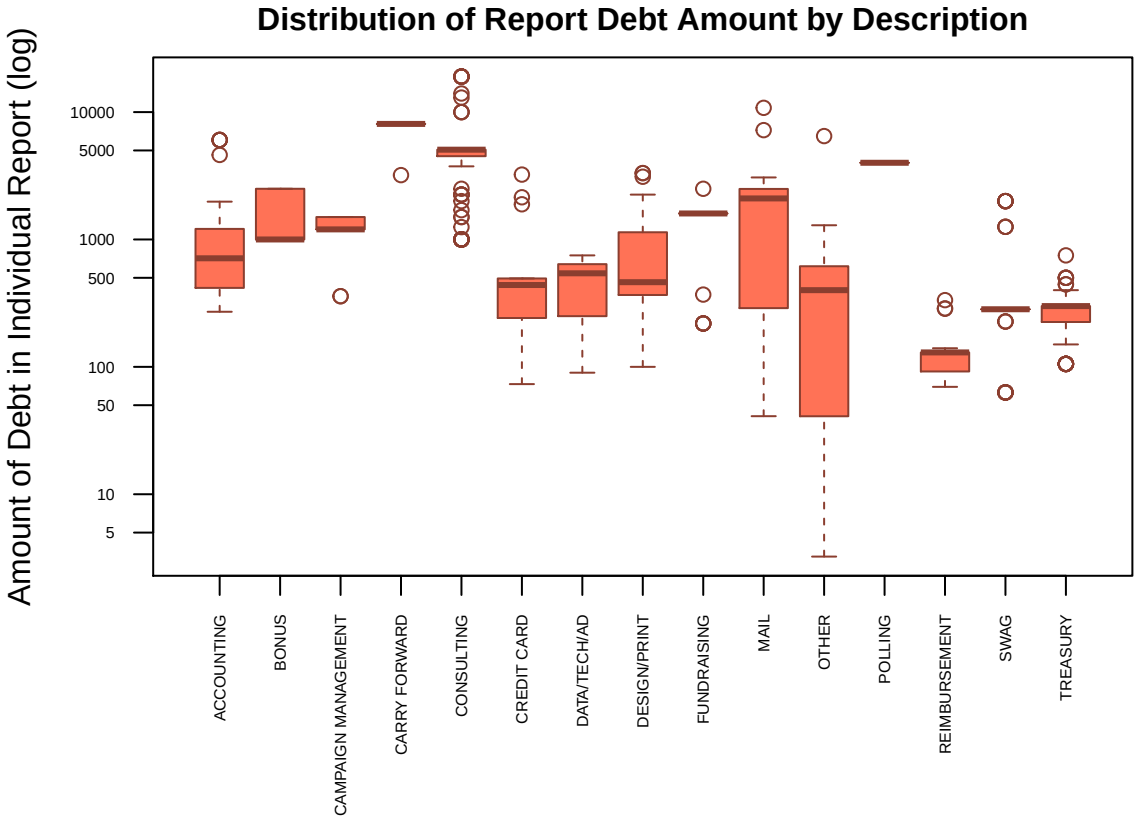
```
      cex.main = 0.7)
```




```
rm(aggr_descr)
```

In order to analyze the distribution of debt amount by description group, we decided to look at it in terms of amount spent. Otherwise, it was hard to see any patterns. The boxplot below shows that consulting expectedly had the highest mean (after carry forward), but it is interesting to note how small Q1-Q3 range is (with a lot of outliers). It is due to the fact that a lot of the reports in this category had exactly the same amount.

```
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)
```



3.d Analysis of Debt by Vendor

Looking back at our univariate analysis, we recall that there were 84 transactions made of $5070.14 and eight of $19,000. It becomes clear that these expenditures were related to a specific vendor consulting service.

```
table(subset(CandidateDebtSub, amount_num == 5070.14)$vendorname)
```

```
##
## HIRSCHBERG STRATEGIES INC.
##                         84
```

```
table(subset(CandidateDebtSub, amount_num == 19000)$vendorname)
```

```
##
## NEW PARTNERS CONSULTING INC.
##                            8
```

```
table(subset(CandidateDebtSub, amount_num == 5070.14)$description_aggr)
```

```
##
## CONSULTING
##         84
```

```
table(subset(CandidateDebtSub, amount_num == 19000)$description_aggr)
```

```
##
## CONSULTING
##          8
```

```
vendor_revenue = aggregate(CandidateDebtSub$amount_num, by=list(vendorname=CandidateDebtSub$vendorname), FUN=sum
head(setNames(vendor_revenue[with(vendor_revenue, order(-x)), ], c('Vendor Name', 'Total Revenue')), 5)
```

```
##                    Vendor Name Total Revenue
## 27   HIRSCHBERG STRATEGIES INC.     425891.76
## 40 NEW PARTNERS CONSULTING INC.     152000.00
## 36               MCINTIRE JAMES     129200.00
```

```
## 53  PROJECT ACCOUNTING SERVICES      94592.75
## 7               ARGO STRATEGIES      91826.36
```

Indeed, we see that these two vendors were the largest revenue generators, as they accounted for 32% and 11% of the total amount reportedly spent by candidates.

4. Analysis of Secondary Effects

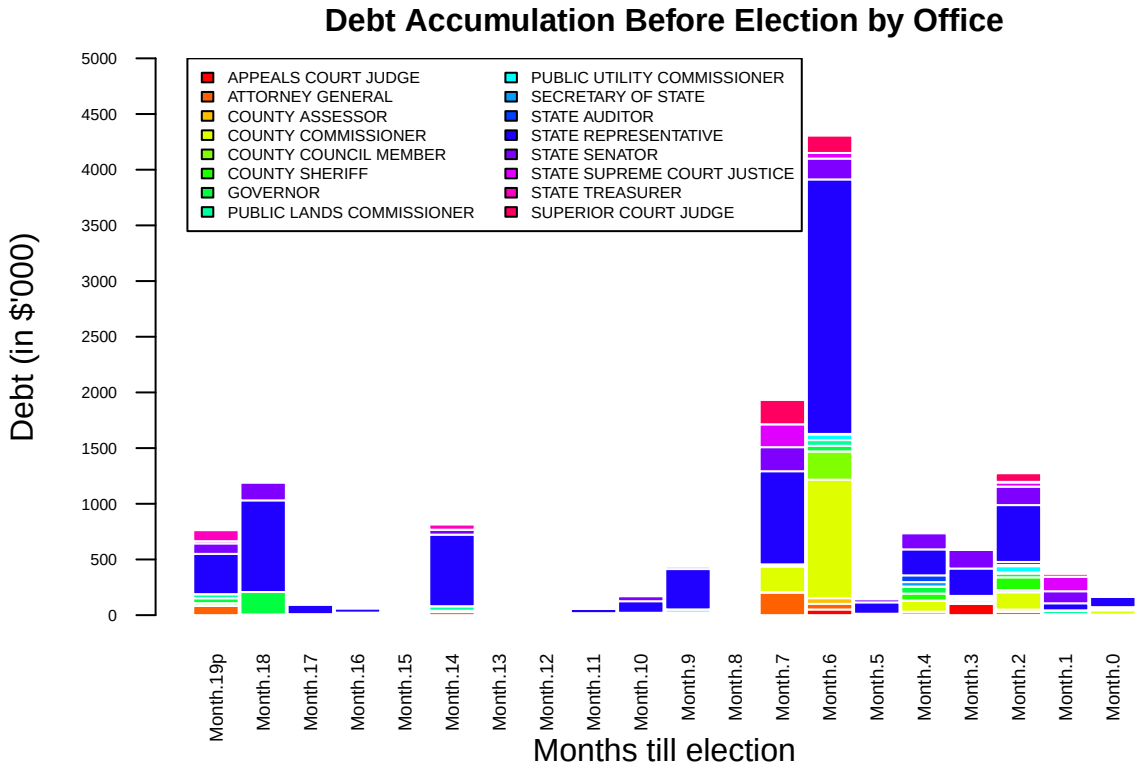
There are multiple secondary effects that come into play while trying to understand candidate debt. It is actually quite challenging to look only at bivariate relationships in this data set. So far, the analysis indicated that debt amounts differ by office, by month they occured in, by type of transactions. There are also factors that we can't actually analyze at this point (party in particular). For the purposes of this section, we will look into types of interactions of debt amount: (1) with office and month and (2) with office and type of transaction.

4.a Timing Analysis of Debt by Office

We looked previously at how debt occured by months leading to the election. Now if we split this by office (see chart below), we see that "State Representative" candidates were spending disproportionately more 7+ months before the elections. They were still spending more than other candidates in months 0-6, but their share was much smaller Questions to investigate: Are campaigns for "State Representatives" longer than for other offices? Why do candidates start spending so much earlier than others?

```
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(6,5,2,1),
    oma = c(0,0,0,0))
barplot(as.matrix(aggr_months_office2),
        border="white",
        space=0.04,
        cex.names = 0.6,
        las = 2,
        cex.axis = 0.8,
        col = rainbow(16),
        axes = FALSE,
        ylim = range(0, 500000),
        xlab = "Months till election",
        ylab = "Debt (in $'000)")
axis(2, at = seq(0, 500000, 50000),
     cex.axis = 0.5,
     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"))
```

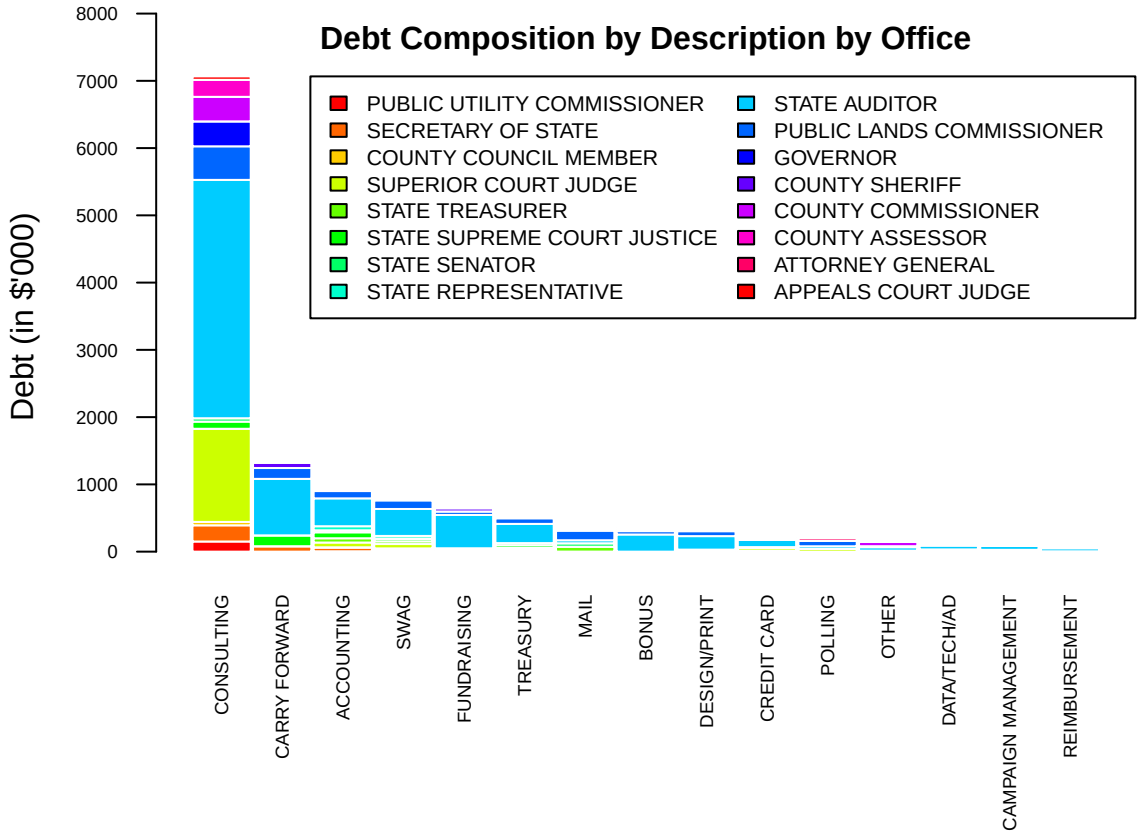
Now we know that "State Representatives" were the largest cohort of candidates, hence, they will dominate absolute debt amounts every month. In the chart below, we analyze time composition of debt for each office. Here we can see that spending patterns by office are very different. "Governor" and "State Treasurer", for example, occure most of their debt 12+ months prior to the election, while candidates for three county offices (commissioner, assessor, and council member) had the largest share of their debts 6 months before the election. Again, we would need to understand office campaign specifics to draw any conclusions.

4.b Composition of Debt by Office by Description

Previously, we identified Consulting as the biggest debt bucket for the candidates. As expected, "State Representative" candidates, as the largest cohort, were responsible for the most spend of this category. "Superior Court Judge" reported the second largest share of consulting-related debt. Other offices also contributed a distinguishable share.

On the other hand, all other debt categories were predominantly generated by "State Representatives". Why is it that debt was more diverse for "State Representatives" compared to other offices? Is it because there were more candidates? Or was it because that campaign was distinctly different from others?

```
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, 800000, 100000),
     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"))
```

5. Conclusion

Several important high level observations can be made about this data:

- Candidates spend the most amout of their money on consulting, though they make frequent swag (ie campagin material) purchases.
- Somewhat ironically, a substantial fraction of the amount spent was spent on the same consultancies, likely for the same services. HIRSCHBERG STRATEGIES INC. and NEW PARTNERS CONSULTING INC. together accounted for over 43% of all money spent. Their services represented relatively high individual expenditures, at $5070.14 and $19,000, respectively
- Spending patterns over the year and full campaign cycle is hard to predict on a month by month basis, however, spending clearly ramps up the year of an election.
- Predictably, the highest number of candidates run for state senate and representative positions.
- The highest spenders per candidate were for public lands commissioner, state treasurer, and appeals court judge, in descending order.
- The offices that were responsible for the highest transaction volume were state representative and county commissioner.
- Reported spending frequency was skewed towards low transaction amounts, while total transaction volume was skewed towards high transaction amounts. Over 60% of transaction volume was from expenditures over $5,000. This demonstrates that a small fraction of the candidates are spending the majority of the money.

While we were able to identify these simple, high level observations, our analysis was significantly constrained by some inconsistencies and shortcomings in the data. In order to improve our analysis in the future, we recommend the following improved practices:

- Force Consistency at Capture Point: Given the multiple unique values per candidate per cyclce for things like 'party,' we recommend a data capture strategy that forces consistency. If a candidate has debt filings with disputing values, we should address the discrepancy directly with the candidate and ask them for clarification.
- Audit of Existing Data: Any other historical data related to candidate debt should be audited for consistency and quality to make sure it doesn't also suffer from the same consistency problems.

- We would like to datasets that extend beyond a single election cycle in order to be able to discern consistencies, discrepancies, and trends between them.
- The dataset should include a comprehensive list of candidates, even those that have not filed debt reports. This data may fill gaps in other variables, such as party affiliation, and point to important patters related to candidates that do and do not comply with these transparency efforts.
- Last, and perhaps most importantly, the data set should provide a significant variable of interest - election outcome. These will enable us to look at how spending and specific vendors influence election outcomes.