Candidate Debt Exploration

Divya Kukati, Jason Tuenge, and Steve Carpenter

25 September 2017

Background

As members of the campaign committee for an upcoming election, we would be interested in knowing as much as possible about debt reported by candidates in previous elections. Using the monthly voter registration statistics for registered voters in Oregon from 2012, we performed an exploratory analysis to help the agency address their goals. Our objectie was to understand how campaign characteristics are related to the candidates debt.

We Were given a dataset on voter registration from 2012 **"CandidateDebt.csv"**. The dependent (or target) variable in this data is named "amount". We were told that the labels of the variables would be as listed below, and that the rest of the variables should be self-explanatory

id: internal identifier that corresponds to a single expenditure record.
reportnumber: 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: Indicates if this record is for a candidate.
filername: The candidate or committee name as reported on the candidates registration.
firstname: This field represents the first name, as reported by the filer.
middleinitial: This field represents the middle initial, as reported by the filer.
lastname: This field represents the last name, as reported by the filer.
office: The office sought by the candidate.
legislativedistrict: The Washington State legislative district.
position: The position associated with an office.
party: The 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.
electionyear: The election year in the case of candidates.
amount: The amount of the debt incurred or order placed.
recordtype: This field designates the item as a debt.
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.
code: The type of debt.
description: The reported description of the transaction.
vendorname: The name of the vendor or recipients 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.

Introduction

The research question that motivated our analysis was: How are campaign characteristics related to the candidates debt?

The Data is in a file called "CandidateDebt.csv" which has 28 variables and 1043 observations. The variable **reportnumber** was of integer class, and the rest were factors with 1-141 levels. All were listed above (i.e., in the assignment), and no 'id' variable was present.

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.1
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

```r
par(oma= c(2,1,5,2) + 0.1) #Sets outside margins : b, l, t, r
par(mar = c(4,1,5,1)) #Sets plot margins
cdebt_original = data.frame(read.csv("CandidateDebt.csv"))
str(cdebt_original)
```

```
## 'data.frame':    1043 obs. of  28 variables:
##  $ reportnumber       : int  100495995 100496548 100498383 100495987 100496259 100496199 100496375 10
##  $ origin             : Factor w/ 1 level "B.3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ filerid            : Factor w/ 141 levels "ASHAK  359","BILLA2 203",..: 110 129 30 122 56 105 93 8
##  $ filertype          : Factor w/ 1 level "Candidate": 1 1 1 1 1 1 1 1 1 1 ...
##  $ filername          : Factor w/ 134 levels "ASHABRANER KARIN L",..: 105 124 31 117 56 99 86 82 70 1
##  $ firstname          : Factor w/ 106 levels "ACHIYAMMA","ALLEN",..: 19 103 43 99 45 72 94 70 64 48
##  $ middleinitial      : Factor w/ 23 levels "","A","B","C",..: 19 15 4 5 18 14 11 19 4 21 ...
##  $ lastname           : Factor w/ 129 levels "ASHABRANER","AXTHELM",..: 101 119 30 113 55 95 82 78 67
##  $ office             : Factor w/ 16 levels "APPEALS COURT JUDGE",..: 12 4 12 6 7 12 4 12 12 12 ...
##  $ legislativedistrict: Factor w/ 15 levels "#N/A","ATTORNEY GENERAL",..: 12 12 12 12 12 12 12 12 12
##  $ position           : Factor w/ 29 levels "","#N/A","1",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ party              : Factor w/ 11 levels "","#N/A","1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ jurisdiction       : Factor w/ 5 levels "#N/A","DEMOCRAT",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ jurisdictioncounty : Factor w/ 52 levels "#N/A","ATTORNEY GENERAL, OFFICE OF",..: 11 11 11 11 11 1
##  $ jurisdictiontype   : Factor w/ 16 levels "","#N/A","BENTON",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ electionyear       : Factor w/ 5 levels "#N/A","Judicial",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ amount             : Factor w/ 2 levels "#N/A","2012": 2 2 2 2 2 2 2 2 2 2 ...
##  $ recordtype         : Factor w/ 124 levels "#N/A","100","1000",..: 50 50 50 50 50 50 50 50 50 50 .
##  $ fromdate           : Factor w/ 2 levels "#N/A","DEBT": 2 2 2 2 2 2 2 2 2 2 ...
##  $ thrudate           : Factor w/ 38 levels "#N/A","01/01/10",..: 17 17 17 17 17 17 17 17 17 17 ...
##  $ debtdate           : Factor w/ 31 levels "#N/A","1/31/10",..: 25 25 25 25 25 25 25 25 25 25 ...
##  $ code               : Factor w/ 73 levels "#N/A","02/01/09",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ description        : Factor w/ 5 levels "","#N/A","Fundraising",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ vendorname         : Factor w/ 107 levels "","#N/A","$750 PER MONTH THROUGH OCTOBER",..: 74 74 74
##  $ vendoraddress      : Factor w/ 76 levels "#N/A","ABBOT TAYLOR",..: 27 27 27 27 27 27 27 27 27 27
##  $ vendorcity         : Factor w/ 81 levels "","#N/A","10 SABLE COURT ",..: 69 69 69 69 69 69 69 69 6
##  $ vendorstate        : Factor w/ 31 levels "","#N/A","BAINBRIDGE ISLAND",..: 31 31 31 31 31 31 31 3
##  $ vendorzip          : Factor w/ 6 levels "","#N/A","CA",..: 6 6 6 6 6 6 6 6 6 6 ...
```

It is already apparent that:

- As we would hope, each **reportnumber** is unique.
- Similarly, there are as many unique **filernames** as unique [**lastname**, **firstname**, **middlename**] triples.
- There are more than 7 **reportnumbers** per **filerid** on average. Consequently, a given person will be counted multiple times if corresponding rows aren't consolidated into a total or set of totals.
- There are 7 more levels for **filerid** than for **filername**. In all cases, simple internet searches revealed that candidates had held one office going into the campaign and ran for a different office for the 2012 election (**filername** in "FROCKT DAVID S", "KELLEY TROY X", "MCKENNA ROBERT M",

"PROBST TIMOTHY P", "WILLIAMS BRENDAN W"), held one office and been appointed another mid-term and by the 2012 election ran for the appointed office (**filername** == "ROLFES CHRISTINE N"), or held no office and started running for one office but ultimately changed their campaign to run for a different office at the end of the election (**filername** == "LADENBURG JOHN W SR").

References:

- https://ballotpedia.org/Dave_Frockt
- https://en.wikipedia.org/wiki/Troy_Kelley
- http://bit.ly/2fnAH9a
- https://en.wikipedia.org/wiki/Rob_McKenna
- https://en.wikipedia.org/wiki/Christine_Rolfes
- https://ballotpedia.org/Tim_Probst
- https://ballotpedia.org/Brendan_Williams

```
length(cdebt_original$reportnumber) - length(unique(cdebt_original$reportnumber))
```

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

```
comb_fmlnames <- length(sort(unique(paste(cdebt_original$lastname,
                                          cdebt_original$firstname,
                                          cdebt_original$middleinitial))))
#shared lastname: FARRELL, JOHNSON, LADENBURG, MORRIS, THOMAS
length(unique(cdebt_original$filername)) - comb_fmlnames
```

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

```
length(unique(cdebt_original$reportnumber)) / length(unique(cdebt_original$filerid))
```

```
## [1] 7.397163
```

```
length(unique(cdebt_original$filerid)) - length(unique(cdebt_original$filername))
```

```
## [1] 7
```

When we reviewed the file we found what seemed to be an undocumented extra column right after the **office** column that was similar in nature but it offset the headers for the rest of the columns. For instance, the **legislativedistrict** data was now under the **position** header, the **party** data was under the **jurisdiction** header, etc… For example, 'DEMOCRAT' should refer to party rather than to jurisdiction. We verified that the **office** column was setup correctly by selecting multiple records and doing an internet search to see what office the specific person (**filername**) was campaigning for in Washington State in 2012. For instance, Sheryl McCloud (**filername** == "MCCLOUD SHERYL G") ran for State Supreme Court Justice and won the seat in 2012. The data in the extraneous columnn indicated that she is somehow linked to a State Representative seat; we couldn't find any evidence of this online.

References:

```
* https://ballotpedia.org/Sheryl_McCloud
* http://sdc.wastateleg.org/frockt/
* https://ballotpedia.org/Washington_elections,_2012
```

We decided to relabel this phantom column to **office2**. In addition, we observed that a number of fields were empty or contained '#N/A' text. We converted these to 'NA' while loading the repaired CSV file, and found that this resulted in several columns (**legislativedistrict**, **position**, **electionyear**) changing from a factor type to an integer type. Similarly, the key **amount** variable changed from a factor type to a numeric type, thereby enabling quantitative analysis of debt.

```
colNames <- c("reportnumber", "origin", "filerid", "filertype", "filername", "firstname",
              "middleinitial","lastname", "office", "office2", "legislativedistrict",
              "position", "party","jurisdiction", "jurisdictioncounty", "jurisdictiontype",
              "electionyear", "amount", "recordtype", "fromdate", "thrudate", "debtdate",
              "code", "description", "vendorname", "vendoraddress", "vendorcity",
              "vendorstate")
cdebt_revised_interim <- read.csv("CandidateDebt.csv", col.names = colNames,
                                  stringsAsFactors = T,
                                  na.strings = c(NA,"NA","#N/A" ,"", " "))
str(cdebt_revised_interim)
```

```
## 'data.frame':    1043 obs. of  28 variables:
##  $ reportnumber       : int  100495995 100496548 100498383 100495987 100496259 100496199 100496375 10
##  $ origin             : Factor w/ 1 level "B.3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ filerid            : Factor w/ 141 levels "ASHAK  359","BILLA2 203",..: 110 129 30 122 56 105 93 8
##  $ filertype          : Factor w/ 1 level "Candidate": 1 1 1 1 1 1 1 1 1 1 ...
##  $ filername          : Factor w/ 134 levels "ASHABRANER KARIN L",..: 105 124 31 117 56 99 86 82 70
##  $ firstname          : Factor w/ 106 levels "ACHIYAMMA","ALLEN",..: 19 103 43 99 45 72 94 70 64 48
##  $ middleinitial      : Factor w/ 22 levels "A","B","C","D",..: 18 14 3 4 17 13 10 18 3 20 ...
##  $ lastname           : Factor w/ 129 levels "ASHABRANER","AXTHELM",..: 101 119 30 113 55 95 82 78 67
##  $ office             : Factor w/ 16 levels "APPEALS COURT JUDGE",..: 12 4 12 6 7 12 4 12 12 12 ...
##  $ office2            : Factor w/ 14 levels "ATTORNEY GENERAL",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ legislativedistrict: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ position           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ party              : Factor w/ 4 levels "DEMOCRAT","INDEPENDENT",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ jurisdiction       : Factor w/ 51 levels "ATTORNEY GENERAL, OFFICE OF",..: 10 10 10 10 10 10 10 10
##  $ jurisdictioncounty : Factor w/ 14 levels "BENTON","CLALLAM",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ jurisdictiontype   : Factor w/ 4 levels "Judicial","Legislative",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ electionyear       : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ amount             : num  283 283 283 283 283 ...
##  $ recordtype         : Factor w/ 1 level "DEBT": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fromdate           : Factor w/ 37 levels "01/01/10","01/01/11",..: 16 16 16 16 16 16 16 16 16 16
##  $ thrudate           : Factor w/ 30 levels "1/31/10","1/31/11",..: 24 24 24 24 24 24 24 24 24 24 ..
##  $ debtdate           : Factor w/ 72 levels "02/01/09","02/01/12",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ code               : Factor w/ 3 levels "Fundraising",..: NA NA NA NA NA NA NA NA NA NA ...
##  $ description        : Factor w/ 105 levels "$750 PER MONTH THROUGH OCTOBER",..: 72 72 72 72 72 72 7
##  $ vendorname         : Factor w/ 75 levels "ABBOT TAYLOR",..: 26 26 26 26 26 26 26 26 26 26 ...
##  $ vendoraddress      : Factor w/ 79 levels "10 SABLE COURT ",..: 67 67 67 67 67 67 67 67 67 67 ...
##  $ vendorcity         : Factor w/ 29 levels "BAINBRIDGE ISLAND",..: 29 29 29 29 29 29 29 29 29 29 ..
##  $ vendorstate        : Factor w/ 4 levels "CA","DC","TX",..: 4 4 4 4 4 4 4 4 4 4 ...
```

We then removed the 56 rows containing NA values in the **amount** column, as these could not be quantitatively related to debt. In fact, these rows were mostly comprised of NA values, thereby confounding comparison with other variables as well.

```
cdebt_revised = subset(cdebt_revised_interim,  !is.na(amount))
#cdebt_revised = subset(cdebt_revised_interim,  !is.na(amount) &
#jurisdiction != "LEG DISTRICT 01 - SENATE")
#str(cdebt_revised)
length(cdebt_original$amount) - length(cdebt_revised$amount)
```

```
## [1] 56
```

A number of variables were still substantially comprised of NA values. As is typical in data analysis, since these values were not impacting our ability to explore the dataset, we decided to keep them in the analysis and handle any evaluations that are impacted by NA values.

```
paste(c(round(1-length(na.omit(cdebt_revised$description))/length(cdebt_revised$description),
              3), "(description)"), collapse = " ")
```

```
## [1] "0.04 (description)"
```

The MM/DD/YY format of the 3 date variables (**fromdate**, **thrudate**, **debtdate**) led to their factor classification. To obtain useful data, We coerced from factor to date.

```
#summary(cdebt_revised$debtdate)
formattedFromdate <- as.Date(as.character(cdebt_revised$fromdate), format = "%m/%d/%y")
formattedThrudate <- as.Date(as.character(cdebt_revised$thrudate), format = "%m/%d/%y")
formattedDebtdate <- as.Date(as.character(cdebt_revised$debtdate), format = "%m/%d/%y")
summary(formattedDebtdate)
```

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

Following are some additional observations made while repairing the CSV file:

- All records associated with **vendorname** == “HICKEY GAYLE” are labeled as **party** == “REPUBLICAN” regardless of the party of the candidate in public records and in the data set for other debt transactions. This has the effect that analysis on party will show records for some candidates as both DEMOCRAT and REPUBLICAN. This is incorrect and should be fixed based on the candidate’s official party alignment. We did some analysis and in no cases did the sample we selected switch from DEMOCRAT to REPUBLICAN or vice versa.
- Both **party** == “INDEPENDENT” records in the data set are also affiliated with a single **vendorname** == “JEFFREYS TERRI M” who just so happens to be a candidate for Mason County Commissioner affiliated as an INDEPENDENT. In some records she is labelled as a DEMOCRAT, in others as a REPUBLICAN (but only in the “HICKEY GAYLE” scenario identified above), and with regard to Postage as an INDEPENDENT. Because the new sources label her as an INDEPENDENT and most of the records label her as a DEMOCRAT, we will not change the party affiliation. The other candidate who used the **vendorname** == “JEFFREYS TERRI M” was a DEMOCRAT named David Frockt. It is assumed he likely shared the cost of postage with Ms Jeffreys and the record was labeled incorrectly.
- Whereas **office** is related to filename, five variables (**office2**, **legislativedistrict**, **jurisdiction**, **jurisdictioncounty**, and **jurisdictiontype**) are all related to one another but do not appear to be related to **office** or **filename**. For example, Terri Jeffreys (http://www.terrijeffreys.org/2012/06/19/citizens-to-elect-terri-jeffreys/) ran for county commissioner (**office**) rather than governor (**office2**), and the corresponding jurisdiction (‘GOVERNOR, OFFICE OF’) would pertain to the latter. The relevance of these 5 variables is unclear.

```
hic <- sum(cdebt_revised$vendorname == "HICKEY GAYLE")
hicrep <- sum((cdebt_revised$vendorname == "HICKEY GAYLE") &
              (cdebt_revised$party == "REPUBLICAN"))
hic - hicrep
```

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

```
sum(cdebt_revised$vendorname == "JEFFREYS TERRI M")
```

```
## [1] 2
```

```
sub_tjven <- subset(cdebt_revised, vendorname ==  "JEFFREYS TERRI M",
                    c(filename, party, vendorname))
sub_tjven
```

```
##             filename       party       vendorname
## 704 JEFFREYS TERRI M INDEPENDENT JEFFREYS TERRI M
## 842   FROCKT DAVID S INDEPENDENT JEFFREYS TERRI M
```

```
sub_tjdfc <- subset(cdebt_revised, ((filername == "JEFFREYS TERRI M") |
                                    (filername == "FROCKT DAVID S")),
                    c(filername, party, vendorname))
sub_tjdfc[order(sub_tjdfc$filername),]
```

```
##            filername       party                 vendorname
## 81    FROCKT DAVID S  REPUBLICAN                HICKEY GAYLE
## 82    FROCKT DAVID S  REPUBLICAN                HICKEY GAYLE
## 188   FROCKT DAVID S  REPUBLICAN                HICKEY GAYLE
## 355   FROCKT DAVID S    DEMOCRAT              MCINTIRE JAMES
## 360   FROCKT DAVID S    DEMOCRAT             ARGO STRATEGIES
## 467   FROCKT DAVID S    DEMOCRAT PROJECT ACCOUNTING SERVICES
## 598   FROCKT DAVID S    DEMOCRAT         RUDERMAN CONSULTING
## 676   FROCKT DAVID S    DEMOCRAT             ARGO STRATEGIES
## 764   FROCKT DAVID S    DEMOCRAT     TRILOGY INTERACTIVE LLC
## 841   FROCKT DAVID S    DEMOCRAT TRAINING ASSOCIATES PACIFIC
## 842   FROCKT DAVID S INDEPENDENT            JEFFREYS TERRI M
## 965   FROCKT DAVID S    DEMOCRAT PROJECT ACCOUNTING SERVICES
## 27  JEFFREYS TERRI M    DEMOCRAT  HIRSCHBERG STRATEGIES INC.
## 85  JEFFREYS TERRI M  REPUBLICAN                HICKEY GAYLE
## 137 JEFFREYS TERRI M  REPUBLICAN                HICKEY GAYLE
## 187 JEFFREYS TERRI M  REPUBLICAN                HICKEY GAYLE
## 639 JEFFREYS TERRI M    DEMOCRAT             ARGO STRATEGIES
## 704 JEFFREYS TERRI M INDEPENDENT            JEFFREYS TERRI M
## 759 JEFFREYS TERRI M    DEMOCRAT         WINPOWER STRATEGIES
## 846 JEFFREYS TERRI M    DEMOCRAT TRAINING ASSOCIATES PACIFIC
## 911 JEFFREYS TERRI M    DEMOCRAT  HIRSCHBERG STRATEGIES INC.
```

```
sub_tjoff <- subset(cdebt_revised, (filername == "JEFFREYS TERRI M"),
                    c(filername, office, office2))
sub_tjoff
```

```
##            filername             office             office2
## 27  JEFFREYS TERRI M COUNTY COMMISSIONER            GOVERNOR
## 85  JEFFREYS TERRI M COUNTY COMMISSIONER       STATE SENATOR
## 137 JEFFREYS TERRI M COUNTY COMMISSIONER       STATE SENATOR
## 187 JEFFREYS TERRI M COUNTY COMMISSIONER       STATE SENATOR
## 639 JEFFREYS TERRI M COUNTY COMMISSIONER STATE REPRESENTATIVE
## 704 JEFFREYS TERRI M COUNTY COMMISSIONER  COUNTY COMMISSIONER
## 759 JEFFREYS TERRI M COUNTY COMMISSIONER STATE REPRESENTATIVE
## 846 JEFFREYS TERRI M COUNTY COMMISSIONER  COUNTY COMMISSIONER
## 911 JEFFREYS TERRI M COUNTY COMMISSIONER            GOVERNOR
```

Last, the dataset contains several variables that appear unlikely to contribute much or any information:

- **firstname**, **middleinitial**, **lastname**: These are redundant to **filername**, as noted above.
- **origin**, **filertype**, **recordtype**: Factor has only one level. Some rows in **recordtype** are NA, but in these cases the entire row is NA.
- **electionyear**: All values are either either 2012 or NA.

```
summary(cdebt_revised$origin)
```

```
## B.3
## 987
```

```
summary(cdebt_revised$filertype)
```

```
## Candidate
##       987
```

```
summary(cdebt_revised$recordtype)
```

```
## DEBT
##  987
```

```
summary(cdebt_revised$electionyear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2012    2012    2012    2012    2012    2012
```

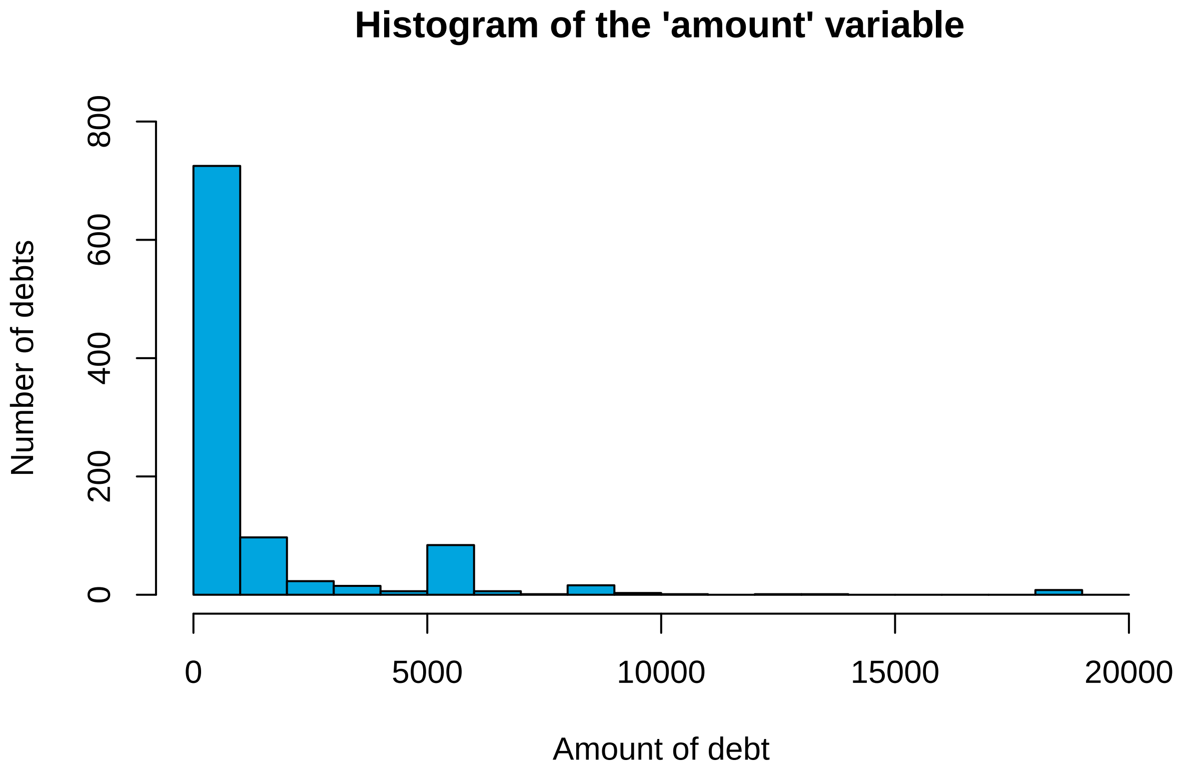
Univariate Analysis of Key Variables

The 'amount' values are all positive and range from around 3 to 19,000 (units are unclear but are presumably dollars or perhaps thousands of dollars), with most values below 1,000. The mean is more than 4x the median, and a histogram confirms the distribution is positively skewed.

```
amount = cdebt_revised$amount
summary(amount)
```

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

```
hist(cdebt_revised$amount, breaks = seq(0, 20000, by = 1000),
     main = "Histogram of the 'amount' variable",  border="black", col="#00A5DF",
     xlab = "Amount of debt" ,ylab="Number of debts", ylim = c(0, 800))
```

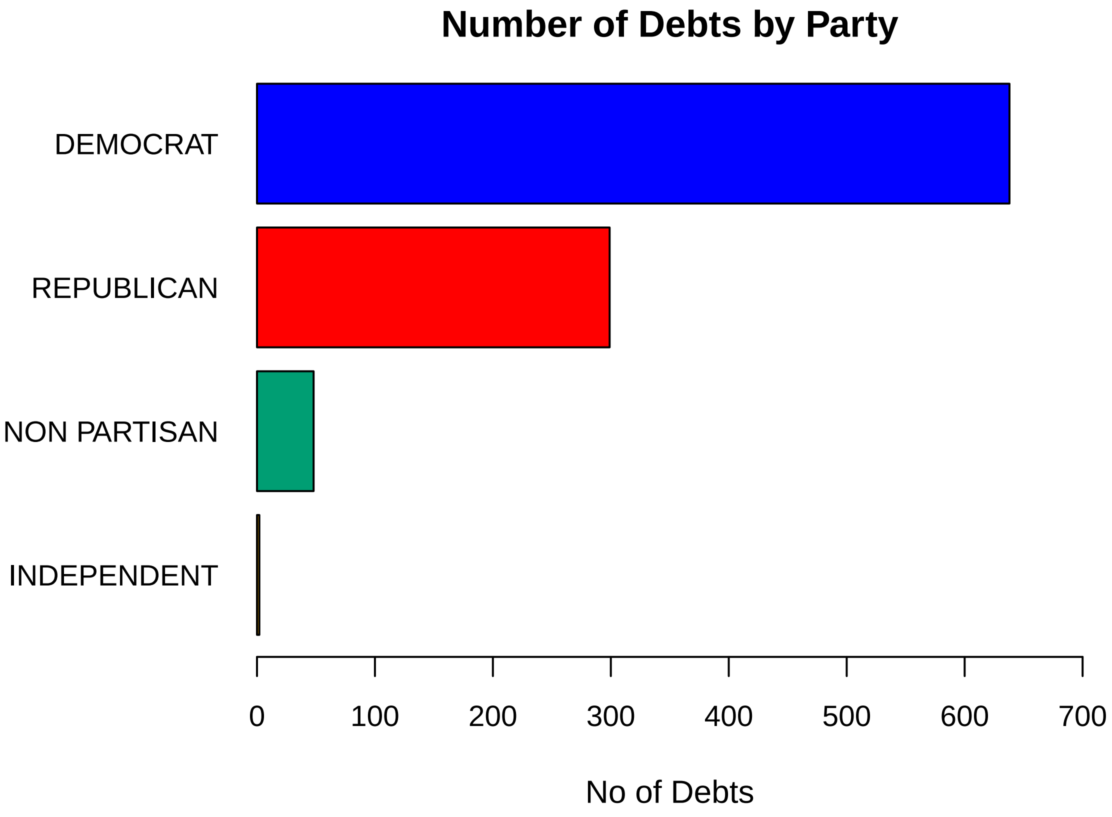


Nearly 65% of filings (by count rather than amount) were by Democrats – more than twice the number filed by Republicans.

```
party <- table(cdebt_revised$party)
round(summary(cdebt_revised$party) / length(cdebt_revised$party), 3)
```

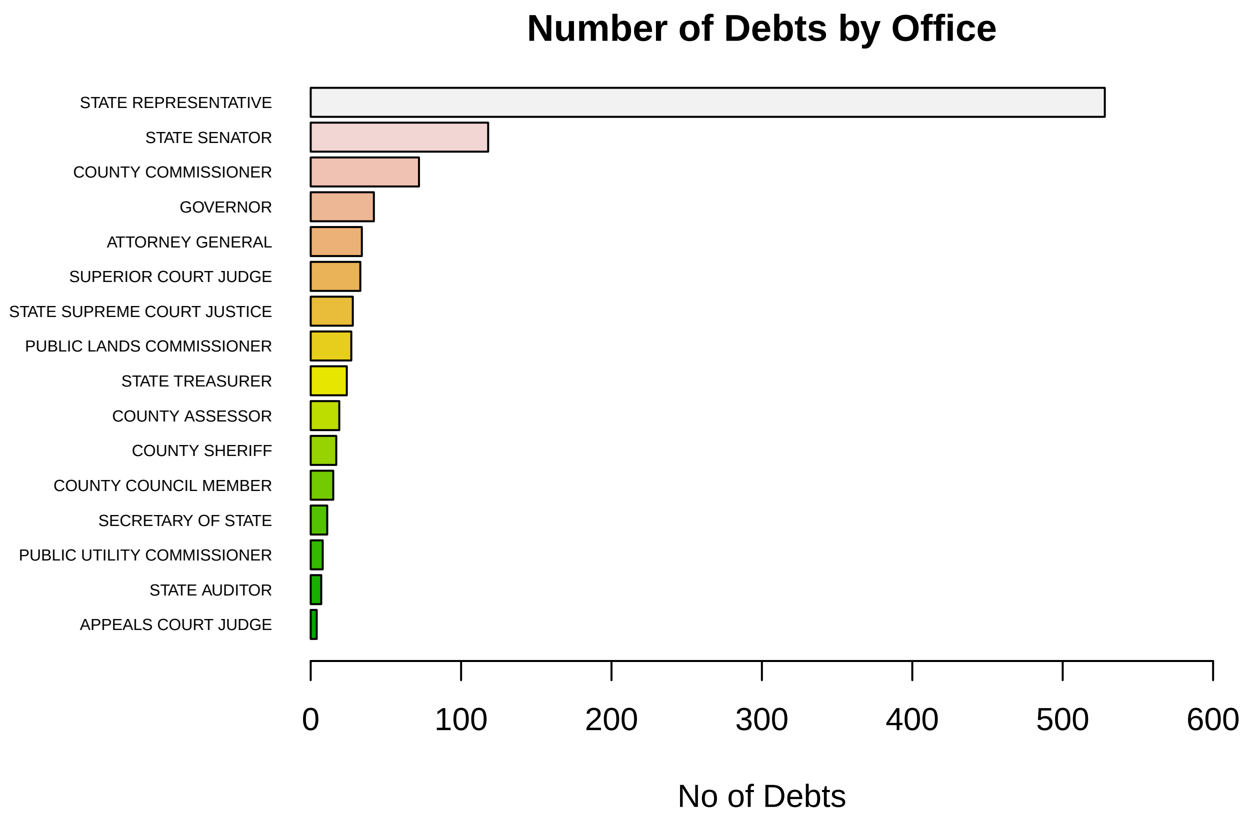
```
##     DEMOCRAT  INDEPENDENT NON PARTISAN   REPUBLICAN
##        0.646        0.002        0.049        0.303
```

```
par(las=1, mar=c(5,10,2,1)) #Sets outside margins : b, l, t, r
colorPalette <- c("#E69F00", "#009E73", "#FF0000", "#0000FF") # I , N , R , D
barplot(party[order(party)], horiz = TRUE , las = 1, main ="Number of Debts by Party" ,
        xlab="No of Debts" ,cex.axis = .9 ,cex.names= 0.9 , col = colorPalette ,
        xlim = c(0, 700))
```



Most debt entries (nearly 54% of total) were filed by state representatives. State senators had the next-largest share, with 12% of the total count.

```
#round(summary(cdebt_revised$office) / length(cdebt_revised$office), 3)
office <- table(cdebt_revised$office)
par(las=1, mar=c(5,8,2, 1)) #Sets plot margins : b, l, t, r
barplot(office[order(office)], horiz = TRUE , las = 1, main ="Number of Debts by Office",
        xlab="No of Debts" ,cex.names=0.5 , col = terrain.colors(16),xlim = c(0, 600))
```

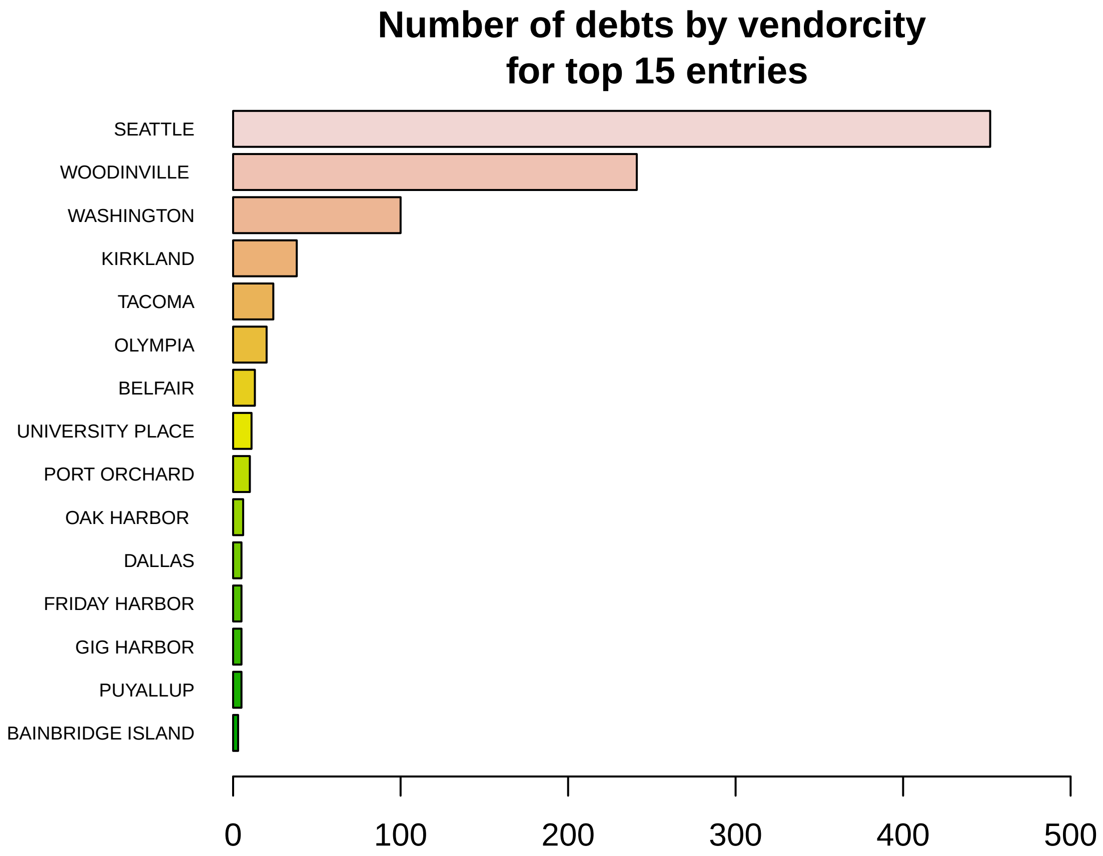


Nearly 47% of entries are associated with vendorcity of Seattle, followed by Woodinville and Washington, D.C.

```
count_Seattle = subset(cdebt_revised, vendorcity == "SEATTLE")
count_cities = subset(cdebt_revised, !is.na(vendorcity))
round(length(count_Seattle$vendorcity) / length(count_cities$vendorcity),3)
```

```
## [1] 0.469
```

```
vendorcity <- table(cdebt_revised$vendorcity)
sortedVendorcity = vendorcity[order(vendorcity, decreasing = TRUE)]
#sortedDescription[15:1] # Top 15 entries
par(las=1, mar=c(2.5,10,2, 0.7)) #Sets plot margins : b, l, t, r
barplot(sortedVendorcity[15:1], horiz = TRUE , las = 1,
        main ="Number of debts by vendorcity\n for top 15 entries",
        xlab="Number of debts" ,cex.names=0.6 , col = terrain.colors(16) ,
        xlim = c(0, 500), width = 10)
```

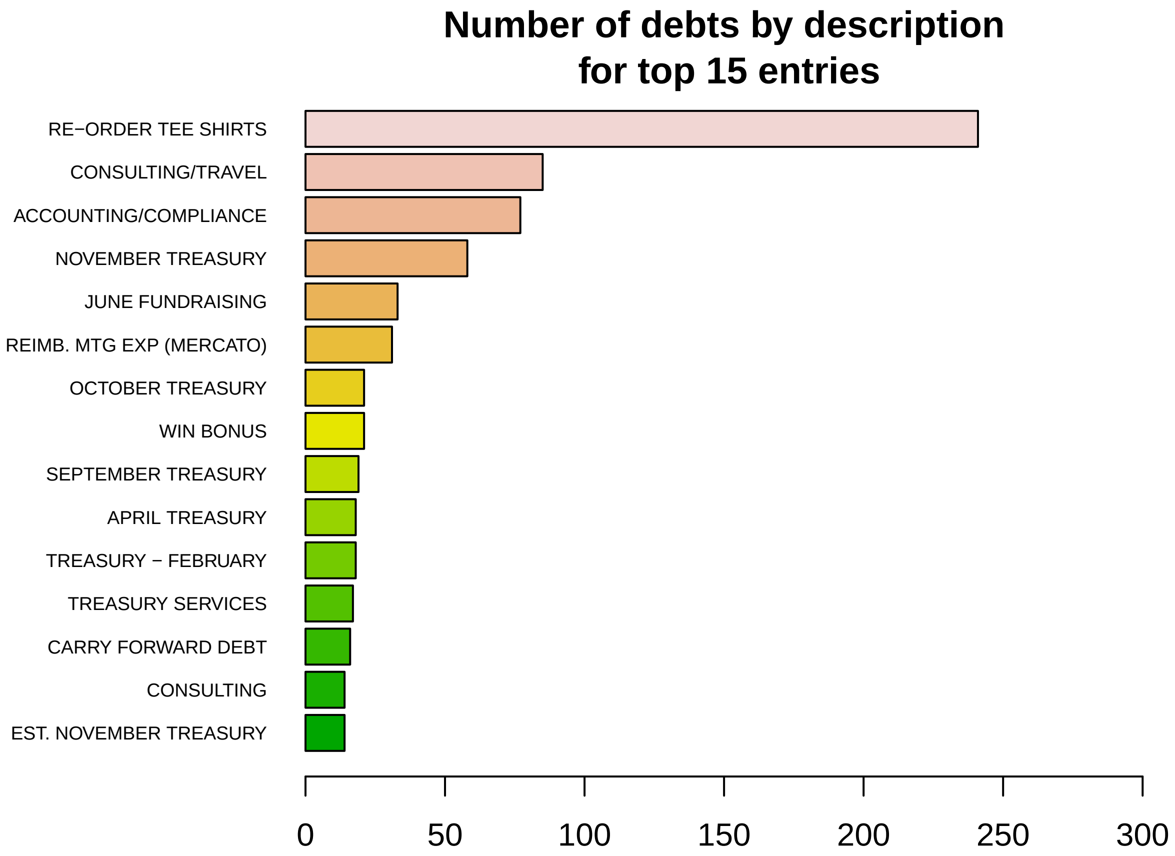


Fully 25% of entries have “RE-ORDER TEE SHIRTS” in the description field. This factor level appears more frequently than the next three levels (CONSULTING/TRAVEL, ACCOUNTING/COMPLIANCE, NOVEMBER TREASURY) combined.

```
count_shirts = subset(cdebt_revised, description == "RE-ORDER TEE SHIRTS")
count_descriptions = subset(cdebt_revised, !is.na(description))
round(length(count_shirts$description) / length(count_descriptions$description),3)
```

```
## [1] 0.254
```

```
description <- table(cdebt_revised$description)
sortedDescription = description[order(description, decreasing = TRUE)]
#sortedDescription[15:1] # Top 15 entries
par(las=1, mar=c(2.5,10,2, 0.7)) #Sets plot margins : b, l, t, r
barplot(sortedDescription[15:1], horiz = TRUE , las = 1,
        main ="Number of debts by description\n for top 15 entries",
        xlab="Number of debts" ,cex.names=0.6 , col = terrain.colors(16) ,
        xlim = c(0, 300) , width = 10)
```

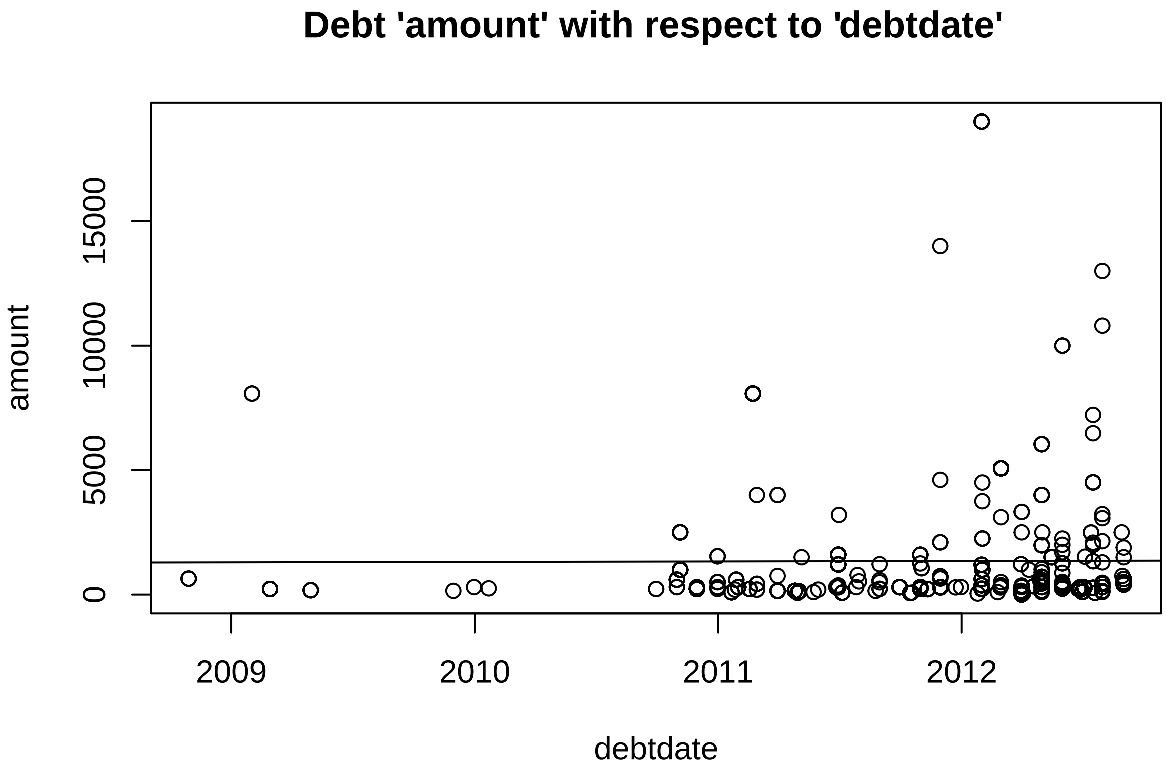


Analysis of Key Relationships

Amount by Debt Date

Although **amount** at first glance appeared to increase with **debtdate**, there was no real correlation. We also observed that there seems to be a surge in debts in the middle of the reportnumber range, presumably due to values for this variable increasing with time (i.e., probably issued sequentially).

```
#plot(cdebt_revised$reportnumber, amount)
#abline(lm(amount ~ cdebt_revised$reportnumber))
plot(formattedDebtdate, amount, main = "Debt 'amount' with respect to 'debtdate'",
     xlab = "debtdate")
abline(lm(amount ~ formattedDebtdate))
```



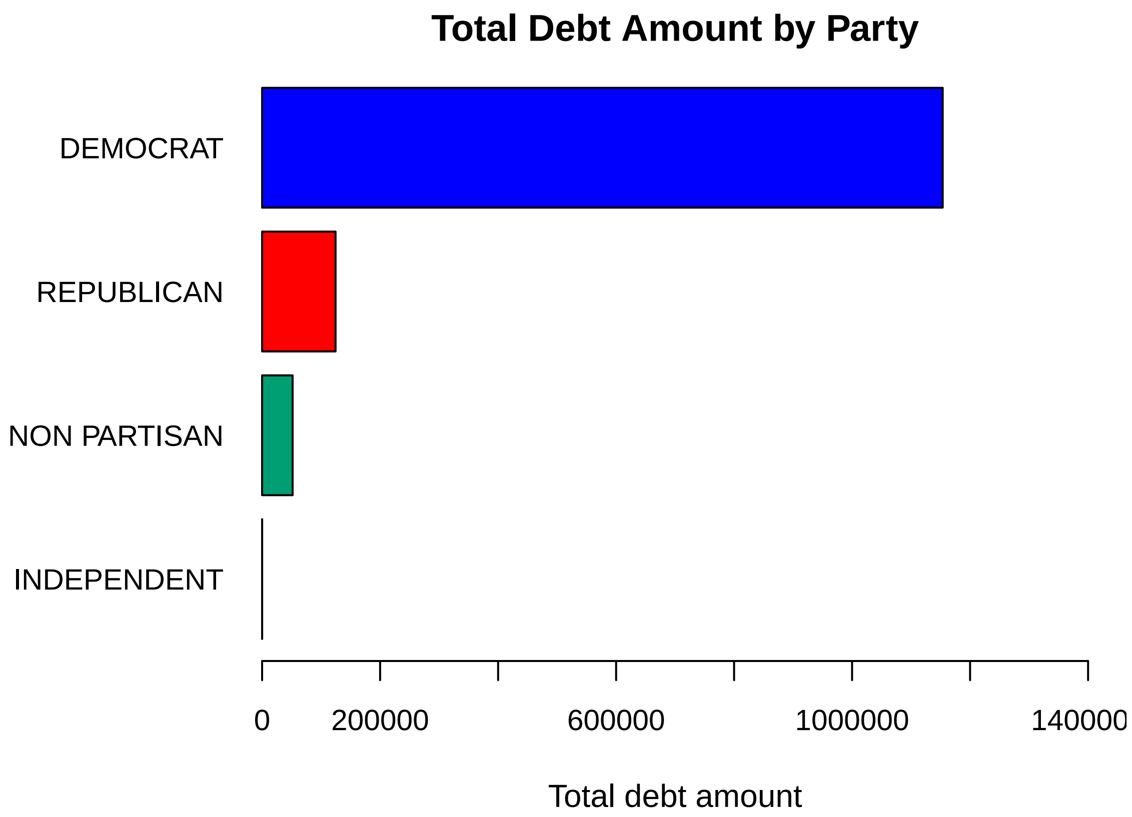
Amount by Party

Most of the debt was filed by Democrats – nearly 10x the amount filed by Republicans. The aforementioned 2x difference in filing count does not explain this gap between parties.

```
amount_sum_party = tapply(cdebt_revised$amount, cdebt_revised$party, FUN=sum )
(amount_sumK_party = amount_sum_party)
```

```
##     DEMOCRAT  INDEPENDENT NON PARTISAN   REPUBLICAN
##   1153600.79       102.88     51696.07    124508.50
```

```
options(scipen=5)
par(las=1, mar=c(5,10,2,1))
colorTAPalette <- c( "#E69F00", "#009E73","#FF0000","#0000FF") # N , I , D , R
barplot(amount_sumK_party[order(amount_sumK_party)], horiz = TRUE , las = 1,
        main ="Total Debt Amount by Party" , xlab="Total debt amount", cex.axis = .9 ,
        cex.names= 0.9 , col = colorTAPalette , xlim=c(0,1400000))
```

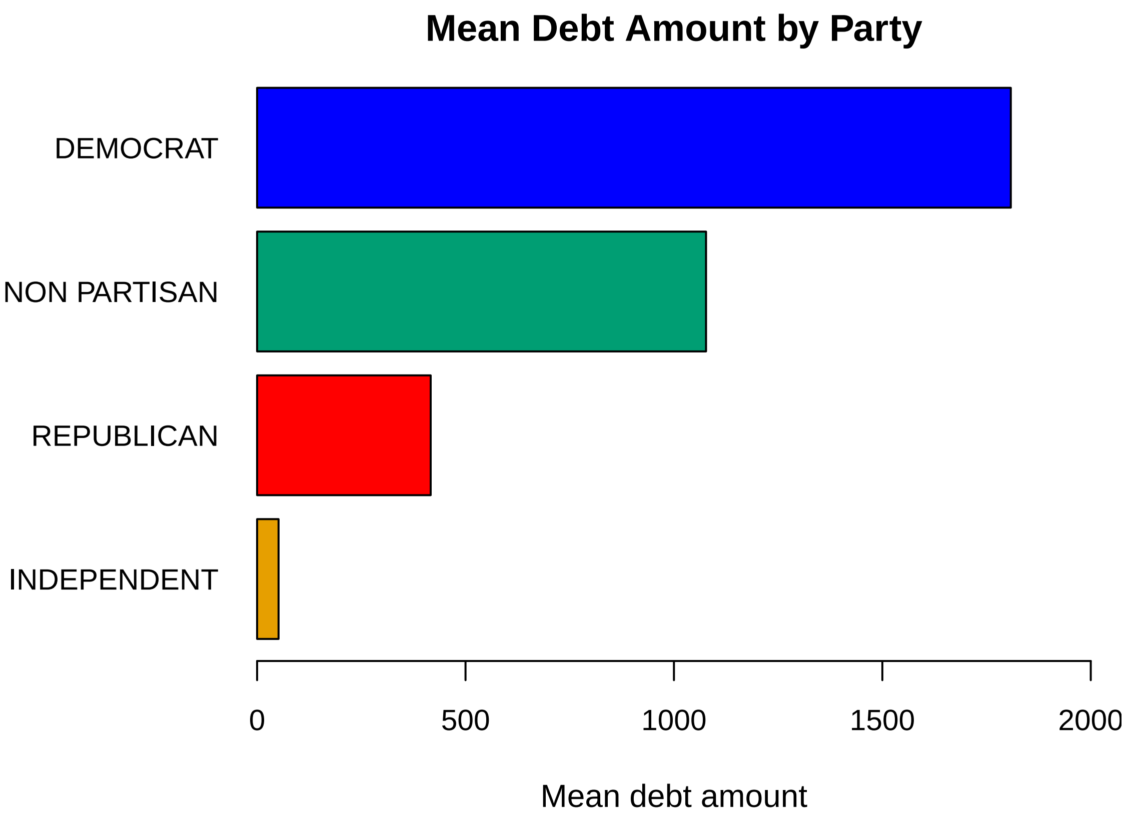


If there were more Democratic candidates than Republican candidates, it wouldn't necessarily be surprising to see Democrats spending more as a party than Republicans. To get a better sense of typical spending spending by candidates of each party, we plot the average.

```
amount_means_party = by(cdebt_revised$amount, cdebt_revised$party,  mean)
amount_means_party[order(amount_means_party, decreasing = TRUE)]
```

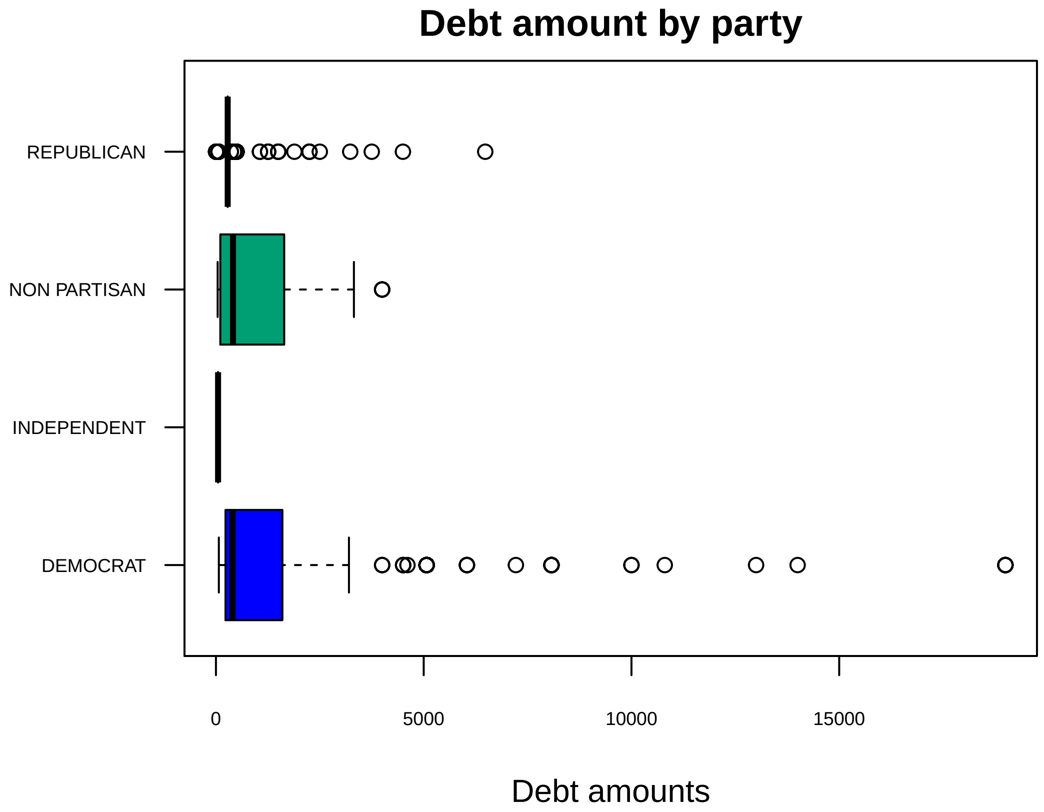
```
## cdebt_revised$party
##     DEMOCRAT NON PARTISAN   REPUBLICAN  INDEPENDENT
##    1808.1517    1077.0015     416.4164      51.4400
```

```
par(las=1, mar=c(5,10,2,0.8)) #Sets outside margins : b, l, t, r
colorMAPalette <- c( "#E69F00", "#FF0000","#009E73","#0000FF") # I , N , R , D
barplot(amount_means_party[order(amount_means_party)], horiz = TRUE , las = 1,
        main ="Mean Debt Amount by Party" ,
        xlab="Mean debt amount" ,cex.axis = .9 ,cex.names= 0.9 ,
        col = colorMAPalette, xlim = c(0, 2000))
```



A boxplot gives a better sense of the distribution, and provides the following insights: * The distribution is quite positively skewed for Democrats, with outliers yielding an average that is more than 4x the median. * Similarly, an outlier results in substantial positive skew for non-partisan candidates. * The numerous 283.25 amounts in the Republican data generate a median of this value, with no interquartile range and multiple "outliers" that yields a relatively mild (in comparison to Democrats) positive skew in the distribution.

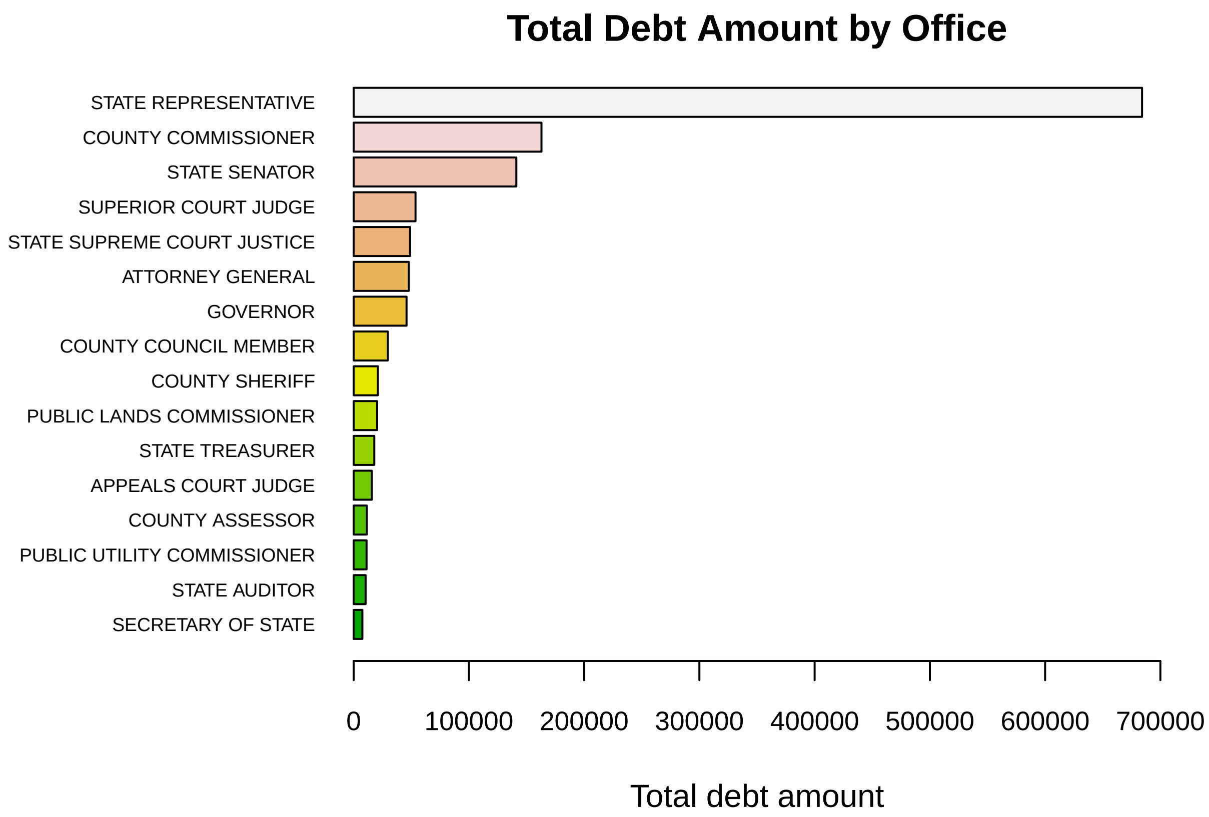
```
#by(cdebt_revised$amount, cdebt_revised$party, median)
par(las=1, mar=c(5,10,2,0.3)) #Sets outside margins : b, l, t, r
colorBAPalette <- c("#0000FF", "#E69F00", "#009E73", "#FF0000" )
# D , R , N , I <==> D , N , I , R
boxplot(cdebt_revised$amount ~ party, cex.axis = .6, horizontal = TRUE ,
        data = cdebt_revised,main = "Debt amount by party",
        col = colorBAPalette, xlab = "Debt amounts")
```



Amount by Office

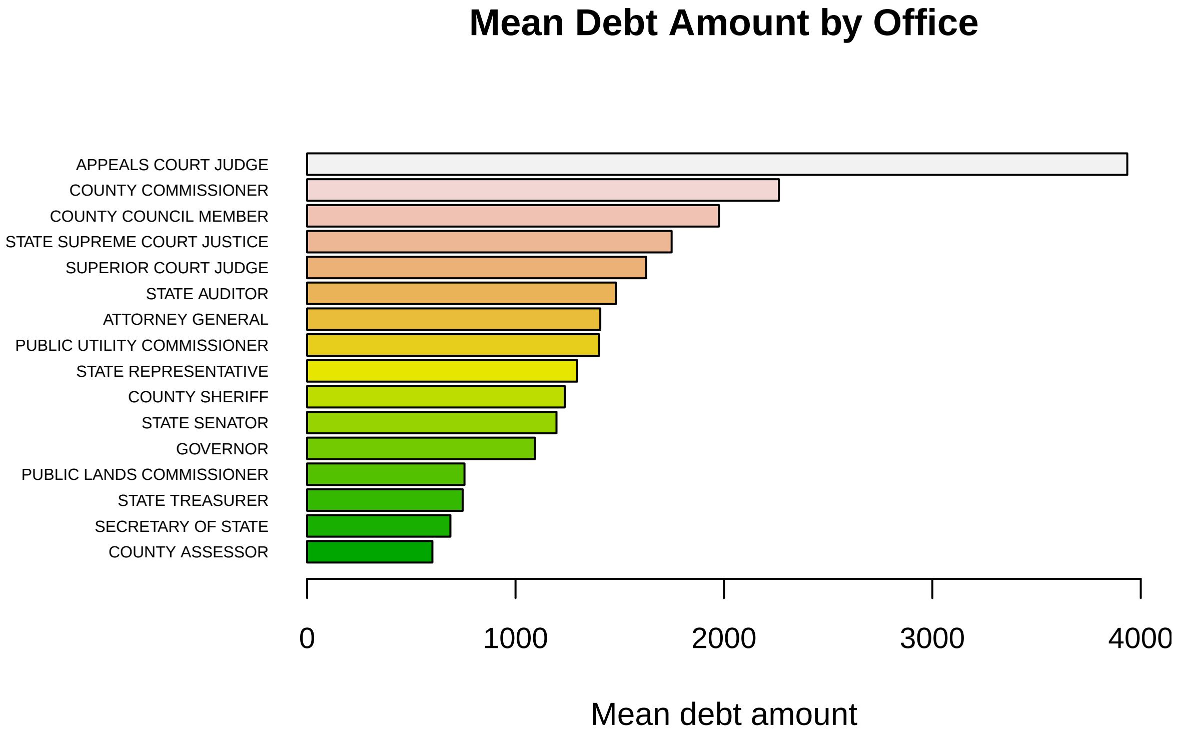
Turning our attention to the relationship between office and debt amount, we see that state representatives filed 4x the amount filed by the second and third-highest offices (county commissioner and state senator, respectively).

```
amount_sumK_office = tapply(cdebt_revised$amount, cdebt_revised$office, FUN=sum)
options(scipen=5)
#amount_sumK_office = amount_sum_office
par(las=1, mar=c(5,10,2,1.5)) #Sets outside margins : b, l, t, r
barplot(amount_sumK_office[order(amount_sumK_office)], horiz = TRUE , las = 1,
        main ="Total Debt Amount by Office" , xlab="Total debt amount" ,cex.axis = .8 ,
        cex.names= 0.6 , col = terrain.colors(16) , xlim=c(0, 700000) )
```



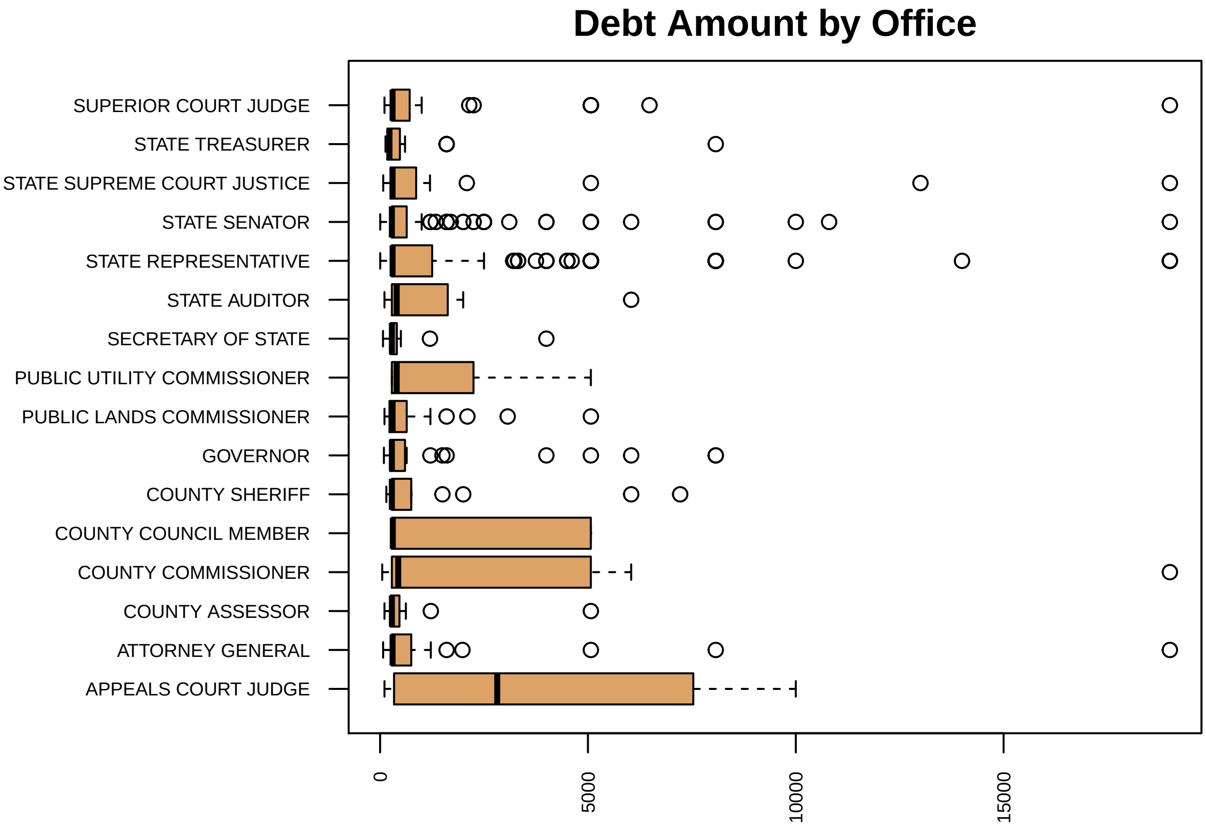
Comparing with mean values as before paints an entirely different picture: We see that appeals court judges appear to have relatively high debt on an individual basis, and state representatives drop toward the lower end of the scale.

```
amount_means_office = by(cdebt_revised$amount, cdebt_revised$office,  mean, na.rm = TRUE)
#amount_means_office[order(amount_means_office, decreasing = TRUE)]
par(las=1, mar=c(5,10,6,0.8)) #Sets outside margins : b, l, t, r
barplot(amount_means_office[order(amount_means_office)], horiz = TRUE , las = 1,
        main ="Mean Debt Amount by Office" , xlab="Mean debt amount" ,cex.axis = .9 ,
        cex.names= 0.5 , col = terrain.colors(16), xlim = c(0, 4000))
```



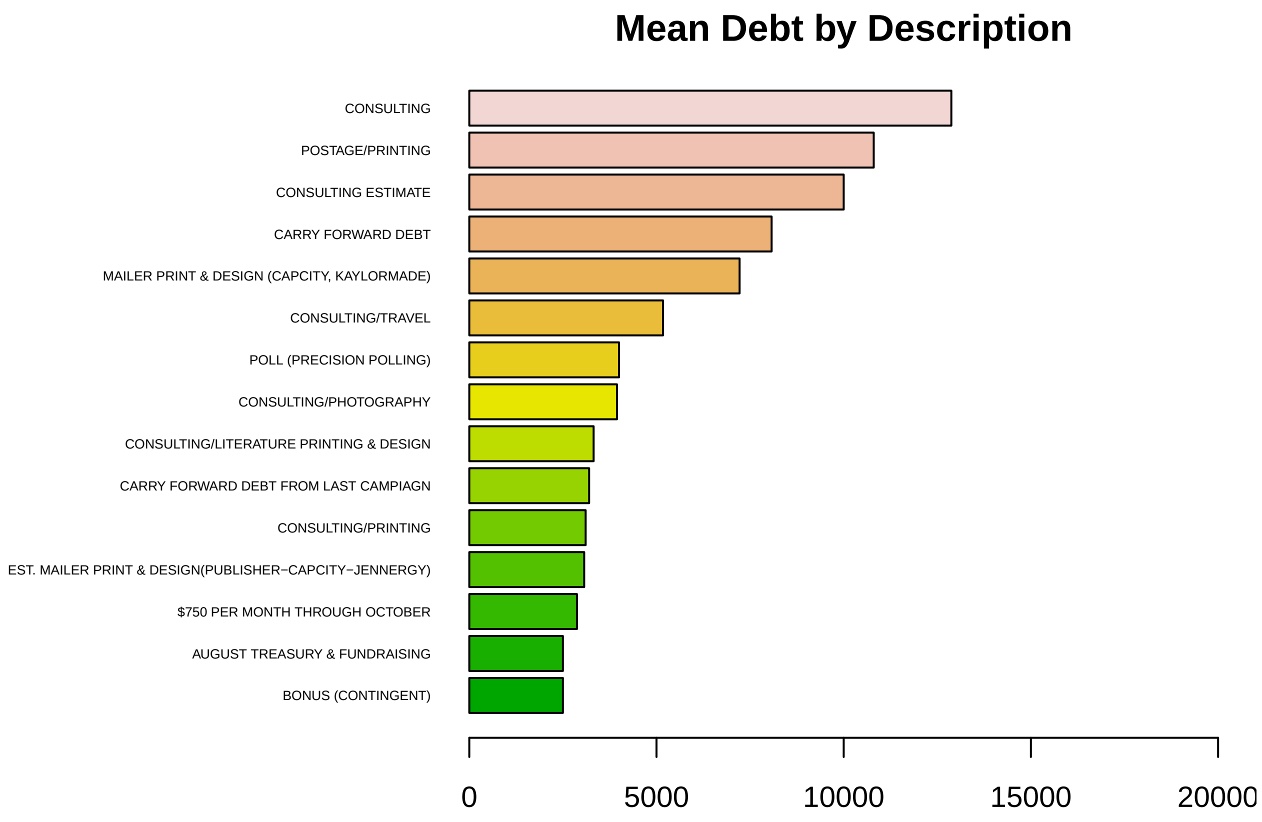
A boxplot again reveals considerable positive skew for all offices excepting appeals court judges, which had no outliers but still exhibited substantial positive skew. Furthermore, the median was around 300 for all offices except appeals court judges, which had a median of nearly 10x this value.

```
#amount_medians_office = by(cdebt_revised$amount, cdebt_revised$office,
#                            median, na.rm = TRUE)
#amount_medians_office[order(amount_medians_office, decreasing = TRUE)]
par(las=1, mar=c(3,10,2,0.3)) #Sets outside margins : b, l, t, r
boxplot(cdebt_revised$amount ~ office, data = cdebt_revised, cex.axis = .6,
        horizontal = TRUE ,
main = "Debt Amount by Office" , las=2 , col = "#DDA266")
```



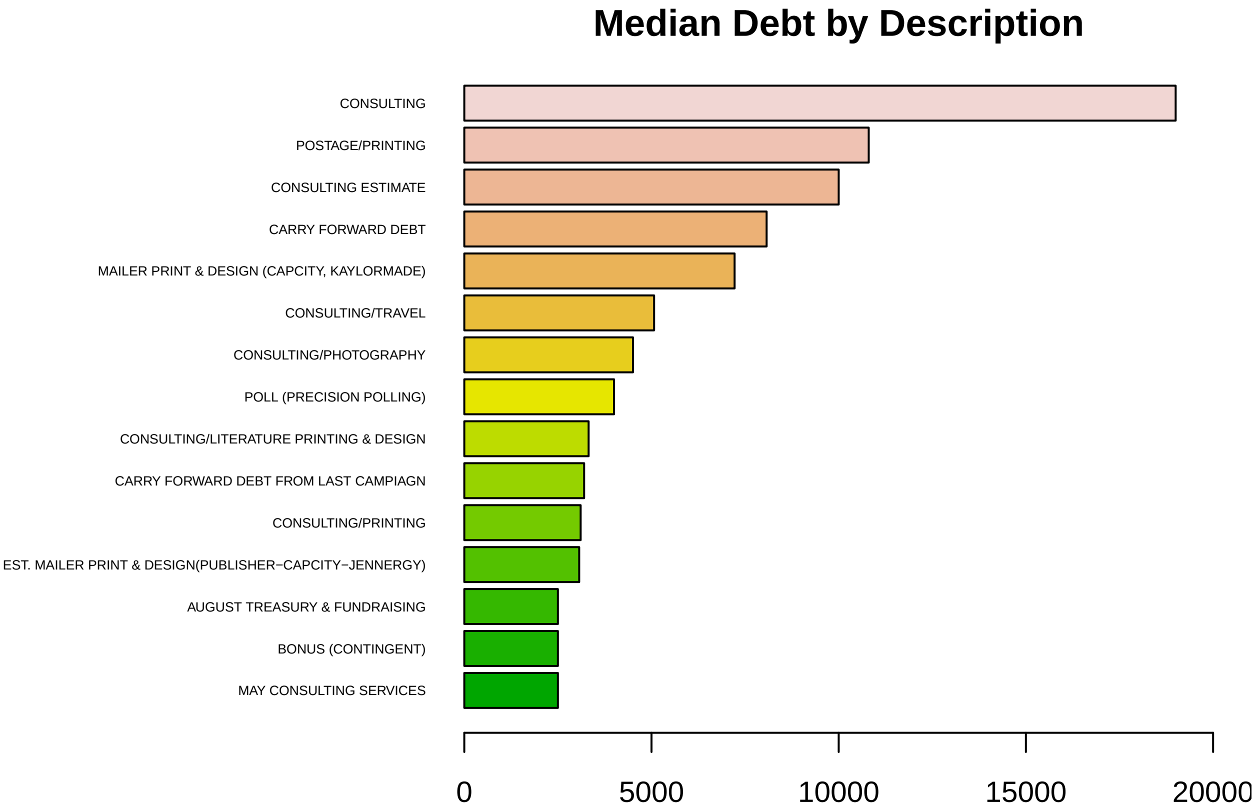
Amount by Description

```
amount_means_description = by(cdebt_revised$amount, cdebt_revised$description,
                              mean, na.rm = TRUE)
sortedMeans = amount_means_description[order(amount_means_description, decreasing = TRUE)]
#sortedMeans[15:1]
par(las=1, mar=c(3,12,2,1))#Sets outside margins : b, l, t, r
barplot(sortedMeans[15:1], horiz = TRUE , las = 1, main ="Mean Debt by Description" ,
        xlab="Mean Debt in $" ,cex.axis = .9 ,cex.names= 0.4 ,
        col = terrain.colors(16), xlim = c(0, 20000))
```



Median for Consulting is pulled away from the mean which indicates that the distribution is negatively skewed.

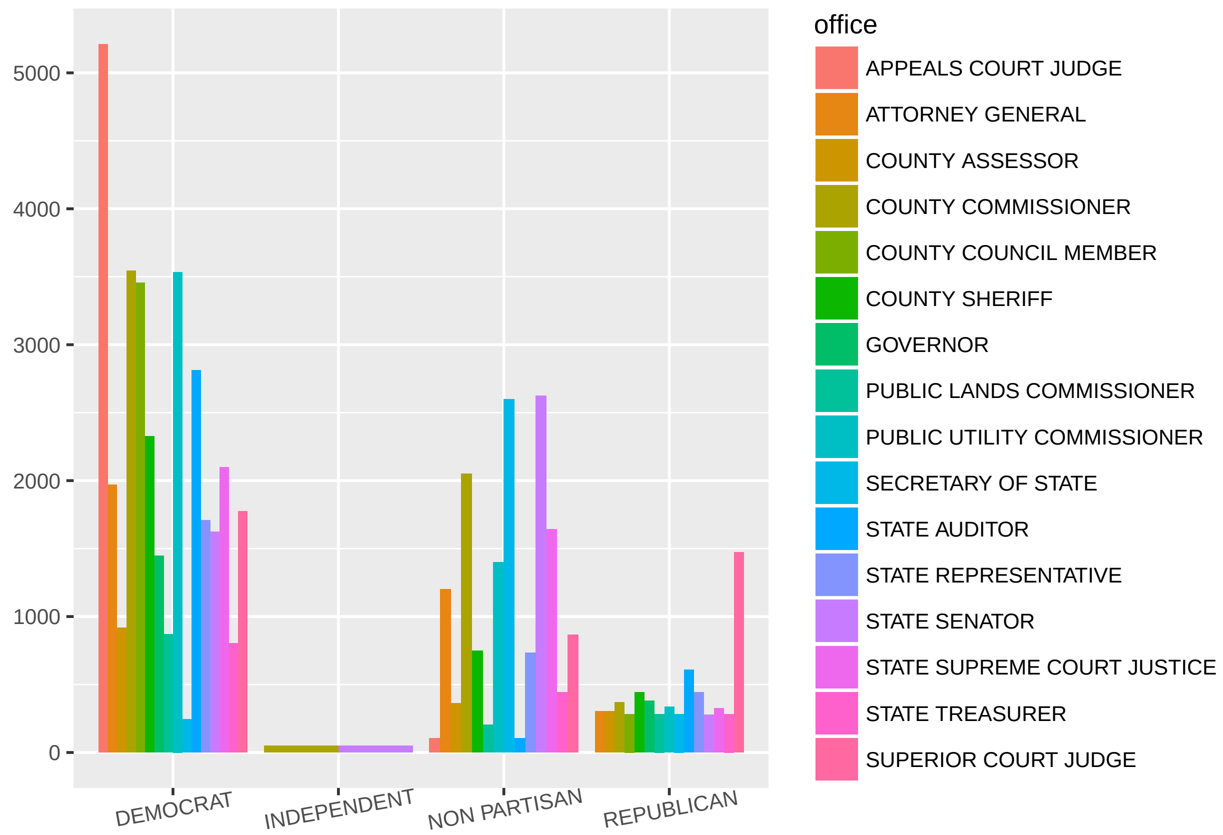
```
amount_median_description = by(cdebt_revised$amount, cdebt_revised$description,
                               median, na.rm = TRUE)
sortedMedian = amount_median_description[order(amount_median_description,
                                               decreasing = TRUE)]
#sortedMedian[15:1]
par(las=1, mar=c(3,12,2,1))#Sets outside margins : b, l, t, r
barplot(sortedMedian[15:1], horiz = TRUE , las = 1, main ="Median Debt by Description" ,
        xlab="Mean Debt in $" ,cex.axis = .9 ,cex.names= 0.4 ,
        col = terrain.colors(16), xlim = c(0, 20000))
```



### Amount, Office, and Party

The plot below is an aggregate of Amount by Office and Party. This plot shows that Candidates the same office but from different parties have different average amount of debts. For instance, Democratic state representatives have average debt of 1710, whereas Republican state representatives have average debt of 443. Non-Partisan state senator have average debt of 2624, whereas Democratic state senators have average debt of 1623, which contadricts the average amount by Office.

```
amount_office_party <- aggregate(x = cdebt_revised$amount,
                                 by = cdebt_revised[c("office", "party")],
                                 FUN = mean, na.rm=TRUE)
#amount_office_party
par(las=1, mar=c(3,10,2,0.3)) #Sets outside margins : b, l, t, r
ggplot(data=amount_office_party, aes(x=party, y=x, xlab="",ylab="", fill=office)) +
  geom_bar(stat="identity", position=position_dodge()) +
  theme(text = element_text(size=10),axis.text.x=element_text(angle=10,hjust=0.5,vjust=1),
        axis.title.x=element_blank(), axis.title.y=element_blank())
```



Analysis of Secondary Effects

We recognize that full analysis would entail consolidating rows for a given candidate, so that for example an average value across candidates would not be influenced by differing row count per candidate. However, this was prevented by our inability to find an algorithm that would reliably resolve the messy data provided (e.g., some candidates appeared to have not in fact switched parties mid-election).

Conclusion

Real-World Data is very messy and can be influenced by:

- data collection processes
- metadata/data dictionary completeness
- willful instructor sabotage
- formatting for an audience/medium
- encoding techniques
- real-life scenarios (candidates changing their minds)
- analysis of originally unintended scenarios

Real-World Data Analysis is more meaningful if you apply domain knowledge to the data set and use whatever sources you have at your disposal (internet, industry knowledge, etc…) to understand what you are seeing; a good example is the inordinately large number of candidates that seemed to switch parties between Democrat and Republican. This happens in real-life, but based on anecdotal evidence (domain kknowledge) it doesn't happen very often. So when it seems like it is happening more than expected, digging in to understand what is really happening is critical (e.g., Gayle Hickey Tee-Shirt Re-orders are all Republican).