Candidate_Debt_EDA

Introduction

Research Question

How might campaign characteristics be related to candidate debt?

Project Overview

To explore this question, we conducted an Exploratory Data Analysis (EDA). The data set we used came from monthly voter registration statistics for registered voters in Oregon during the 2012 election cycle.

Data Caveats/Limitations

1) This data set was limited to state and local elections in Oregon, so results from this analysis should not be assumed to be generalizable beyond this.
2) As the data set is limited to only one election term (2012), we cannot determine debt held over longer periods of time. In addition to this, this data does not provide insight into candidate behavior and patterns over time.
3) It is reasonable to assume that an incumbent might spend less and/or go less into debt than a challenger. However, this information was not included in the data set. Being able to hold incumbency constant, even, might provide a deeper understanding of the truly significant and relevant variables.
4) While this EDA explores relationships between variables, we cannot infer causality.

Data Loading and Cleaning

Before loading our data set, we loaded the libraries necessary for our analysis. We also use a number of parameters that allowed us to quickly turn on/off viewing the data tables we created through our EDA.

```
library(readr)
library(moments)
library(plyr)
library(dplyr)
library(tidyr)
library(car)

save_transformed_datatables=T

show_corrupted_file=F
show_uncorrupted_file=F
show_clean_file=F
show_with_and_without_na=F
show_transformed_datatable=F
```

We set the working directory before loading the data using the import csv function and handled the 'NA' values using the read.table command.

```
my_path="./"
setwd(my_path)
getwd()
```

```
## [1] "/Users/marycboardman/Dropbox/Data_Science/Portfolio/Candidate-Debt"
```

The file included a header row, but we soon realized there was corruption in the file header due to the mis-alignment of column headers to the data. We therefore loaded the file and using the code below added an unknown column header for the unknown categorical column in the dataset. After this we persisted a new file leaving the original file intact should we need to address further corruption issues.

```
csv_seperator = ","

s <- readLines("./CandidateDebt.csv")
no_lines = length(s)

if (show_corrupted_file)
{
  View(s)
}

old_header = s[[1]]
old_header_split <- strsplit(s[[1]],",")

no_columns = lengths(old_header_split)-1
new_header_split <- c(old_header_split[[1]][0:11],"unknown",old_header_split[[1]][12:no_columns])

(s[[1]] <- paste(new_header_split,sep = "",collapse=","))
```

```
## [1] "reportnumber,origin,filerid,filertype,filername,firstname,middleinitial,lastname,office,legisla
```

```
writeLines(s, "./CandidateDebt_clean.csv")

if (show_uncorrupted_file)
{
  View(s)
}
```

We then took the clean file and read it into *R* in its un-modified form. It was observed that the read.table operation did not cast the columns as we wanted and therefore we would need to do more work to format it correctly.

```
candidate_debt <- read.table("./CandidateDebt_clean.csv",
                             na.strings=c("NA"), as.is = T ,
                             header = TRUE,
                             sep = ",",
                             quote = "\"",
                             dec = ".",
                             fill = TRUE,
                             comment.char = "")

if (show_clean_file==T)
{
    View(head(candidate_debt))
}
```

Initially, we checked the row count of the raw imported data file and found 1043 rows, representing 1043

possible observations. We later realized that some were 'NA', as shown below.

```
(imported_rows = nrow(candidate_debt))
```

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

Then, we checked for 'NA' values. To do this, we identified each row in the dataset where any columns contain a row that includes an #N/A in any of the cells. The filter vector includes the row number of each row that matches #N/A.

```
filter <- unique (unlist (lapply (candidate_debt, function (x) which (x=="#N/A"))))
```

We then extracted two dataframes, one that includes the rows with an #N/A and a second dataframe from which they have been removed. We also included a parameterized call to enable us to turn on viewing of these dataframes using the view command.

```
candidate_debt_clean <- candidate_debt[-(filter),,drop=FALSE]
candidate_debt_bad <- candidate_debt[filter,,drop=FALSE]

if (show_with_and_without_na==T)
{
    View(head(candidate_debt_clean))
    View(candidate_debt_bad)
}
```

Then, we checked the row count of each data frame to ensure it matched the row count of the file we initially loaded.

```
(clean_rows <- nrow(candidate_debt_clean))
```

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

```
(bad_rows <- nrow(candidate_debt_bad))
```

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

```
assertthat::are_equal(imported_rows,clean_rows+bad_rows)
```

```
## [1] TRUE
```

This showed actual observations, once the 'NAs' were accounted for and cleaned from the data set.

Next, we transformed, cast and parsed rows to each of the applicable data types and cast specific columns to factors.

We set the vendorzip column to a zero length string, as it is completely empty in the imported file. This way, we allowed for the use of NA checks for later analysis. Otherwise, they might impact our results when we want to use some of the na.action property in some of the charting and analysis.

Then, we cast dates from character strings to dates using the format m/d/yyyy in order to make it more intuitive and user friendly to work with.

```
candidate_debt_transformed <- transform(candidate_debt_clean,
      amount = as.numeric(amount),
      fromdate = as.Date(fromdate, format = "%m/%d/%Y"),
      debtdate = as.Date(debtdate, format = "%m/%d/%Y"),
      thrudate = as.Date(thrudate, format = "%m/%d/%Y"),
      filertype = as.factor(filertype),
      candidate = filername,
      office = as.factor(office),
      legislativedistrict = as.factor(legislativedistrict),
```

```
      position = as.factor(position),
      party = as.factor(party),
      jurisdiction = as.factor(jurisdiction),
      jurisdictioncounty = as.factor(jurisdictioncounty),
      jurisdictiontype = as.factor(jurisdictiontype),
      electionyear = as.factor(electionyear),
      code = as.factor(code),
      recordtype = as.factor(recordtype),
      vendorstate = as.factor(vendorstate),
      vendorzip = "")

if (show_transformed_datatable==T)
{
    View(head(candidate_debt_transformed))
}
```

Another issue that we found is that the same candidates often list different parties. Given how often this happens in the data, it casts doubt as to the reliability of the information about party.

```
party_cand = by(candidate_debt_transformed$party, candidate_debt_transformed$filername, summary)
head(party_cand, n = 1)
```

```
## $`ASHABRANER KARIN L`
##     DEMOCRAT  INDEPENDENT NON PARTISAN   REPUBLICAN
##            7            0            0            2
```

Description of Data

Below is our the documented code book of the imported file. It shows the number of observations, as well as the variables, variable types, and summary statistics for non-categorical variables. These summary statistics include mean, median, along with 1st and 3rd quartiles.

```
names(candidate_debt_transformed)
```

```
##  [1] "reportnumber"        "origin"              "filerid"
##  [4] "filertype"           "filername"           "firstname"
##  [7] "middleinitial"       "lastname"            "office"
## [10] "legislativedistrict" "position"            "unknown"
## [13] "party"               "jurisdiction"        "jurisdictioncounty"
## [16] "jurisdictiontype"    "electionyear"        "amount"
## [19] "recordtype"          "fromdate"            "thrudate"
## [22] "debtdate"            "code"                "description"
## [25] "vendorname"          "vendoraddress"       "vendorcity"
## [28] "vendorstate"         "candidate"           "vendorzip"
```

```
summary(candidate_debt_transformed)
```

```
##   reportnumber          origin            filerid             filertype
##  Min.   :100346104   Length:987         Length:987         Candidate:987
##  1st Qu.:100446276   Class :character   Class :character
##  Median :100471547   Mode  :character   Mode  :character
##  Mean   :100466089
##  3rd Qu.:100494036
##  Max.   :100599472
##
```

```
##   filername          firstname         middleinitial
##  Length:987         Length:987         Length:987
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    lastname                          office
##  Length:987         STATE REPRESENTATIVE:528
##  Class :character   STATE SENATOR       :118
##  Mode  :character   COUNTY COMMISSIONER : 72
##                     GOVERNOR            : 42
##                     ATTORNEY GENERAL    : 34
##                     SUPERIOR COURT JUDGE: 33
##                     (Other)             :160
##            legislativedistrict    position     unknown
##  STATE REPRESENTATIVE:384             :298   Length:987
##  STATE SENATOR       :305      1      :246   Class :character
##  GOVERNOR            :101      43     : 72   Mode  :character
##  ATTORNEY GENERAL    : 49      45     : 68
##  COUNTY COMMISSIONER : 44      21     : 55
##  STATE TREASURER     : 28      11     : 43
##  (Other)             : 76      (Other):205
##           party                             jurisdiction jurisdictioncounty
##  DEMOCRAT    :638   LEG DISTRICT 01 - SENATE   :243   KING     :544
##  INDEPENDENT :  2   GOVERNOR, OFFICE OF        :101            :215
##  NON PARTISAN: 48   LEG DISTRICT 43 - HOUSE    : 72   PIERCE   : 76
##  REPUBLICAN  :299   LEG DISTRICT 45 - HOUSE    : 68   SNOHOMISH: 55
##                     LEG DISTRICT 21 - HOUSE    : 55   SKAGIT   : 34
##                     ATTORNEY GENERAL, OFFICE OF: 49   CLALLAM  : 15
##                     (Other)                    :399   (Other)  : 48
##     jurisdictiontype electionyear     amount          recordtype
##  Judicial   : 36     2012:987     Min.   :    3.24   DEBT:987
##  Legislative:689                  1st Qu.:  283.25
##  Local      : 59                  Median :  300.00
##  Statewide  :203                  Mean   : 1347.42
##                                   3rd Qu.: 1210.50
##                                   Max.   :19000.00
##
##     fromdate             thrudate             debtdate
##  Min.   :0009-10-01   Min.   :0009-10-31   Min.   :0008-10-29
##  1st Qu.:0011-10-01   1st Qu.:0011-10-31   1st Qu.:0011-07-03
##  Median :0012-02-01   Median :0012-02-29   Median :0012-02-29
##  Mean   :0011-12-19   Mean   :0012-01-20   Mean   :0011-12-13
##  3rd Qu.:0012-06-01   3rd Qu.:0012-07-16   3rd Qu.:0012-07-03
##  Max.   :0012-08-01   Max.   :0012-08-31   Max.   :0012-08-31
##
##                      code     description         vendorname
##                        :610   Length:987         Length:987
##  Fundraising           :  5   Class :character   Class :character
##  Management Services   : 10   Mode  :character   Mode  :character
##  Operation and Overhead:362
##
```

```
##
##
##  vendoraddress       vendorcity        vendorstate             candidate
##  Length:987         Length:987           : 25      GOLDMARK PETER J : 27
##  Class :character   Class :character   CA: 10      MCINTIRE JAMES L : 24
##  Mode  :character   Mode  :character   DC:100      BROWN LISA J     : 23
##                                        TX:  5      CHOPP FRANK V    : 23
##                                        WA:847      FERGUSON ROBERT W: 23
##                                                    INSLEE JAY R     : 20
##                                                    (Other)          :847
##  vendorzip
##  :987
##
##
##
##
##
##
```

```
str(candidate_debt_transformed)
```

```
## 'data.frame':    987 obs. of  30 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          : Factor w/ 1 level "Candidate": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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             : Factor w/ 16 levels "APPEALS COURT JUDGE",..: 12 4 12 6 7 12 4 12 12 12 ...
##  $ legislativedistrict: Factor w/ 14 levels "ATTORNEY GENERAL",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ position           : Factor w/ 28 levels "","1","11","14",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ unknown            : chr  "" "" "" "" ...
##  $ 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/ 15 levels "","BENTON","CLALLAM",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ jurisdictiontype   : Factor w/ 4 levels "Judicial","Legislative",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ electionyear       : Factor w/ 1 level "2012": 1 1 1 1 1 1 1 1 1 1 ...
##  $ amount             : num  283 283 283 283 283 ...
##  $ recordtype         : Factor w/ 1 level "DEBT": 1 1 1 1 1 1 1 1 1 1 ...
##  $ fromdate           : Date, format: "0012-06-01" "0012-06-01" ...
##  $ thrudate           : Date, format: "0012-07-16" "0012-07-16" ...
##  $ debtdate           : Date, format: "0012-07-03" "0012-07-03" ...
##  $ code               : Factor w/ 4 levels "","Fundraising",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ description        : chr  "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-ORDER TEE SHIRTS" "RE-O
##  $ vendorname         : chr  "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" "HICKEY GAYLE" ...
##  $ vendoraddress      : chr  "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" "PO BOX 2749" ...
##  $ vendorcity         : chr  "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " "WOODINVILLE " ...
##  $ vendorstate        : Factor w/ 5 levels "","CA","DC","TX",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ candidate          : Factor w/ 133 levels "ASHABRANER KARIN L",..: 104 123 30 116 55 98 85 81 69 1
##  $ vendorzip          : Factor w/ 1 level "": 1 1 1 1 1 1 1 1 1 1 ...
```

Below are the top rows in our transformed dataframe.

```
head(candidate_debt_transformed)
```

```
##   reportnumber origin    filerid filertype           filername firstname
## 1    100495995    B.3 RYU C  133 Candidate         RYU CINDY S     CINDY
## 2    100496548    B.3 THOMT  368 Candidate THOMAS TIMOTHY N JR   TIMOTHY
## 3    100498383    B.3 FEY J  422 Candidate         FEY JACOB C     JACOB
## 4    100495987    B.3 STRAS  111 Candidate   STRACHAN STEVEN D    STEVEN
## 5    100496259    B.3 INSLJ  110 Candidate        INSLEE JAY R       JAY
## 6    100496199    B.3 RICCM  210 Candidate   RICCELLI MARCUS M    MARCUS
##   middleinitial lastname              office legislativedistrict position
## 1             S      RYU STATE REPRESENTATIVE       STATE SENATOR        1
## 2             N   THOMAS  COUNTY COMMISSIONER       STATE SENATOR        1
## 3             C      FEY STATE REPRESENTATIVE       STATE SENATOR        1
## 4             D STRACHAN       COUNTY SHERIFF       STATE SENATOR        1
## 5             R   INSLEE             GOVERNOR       STATE SENATOR        1
## 6             M RICCELLI STATE REPRESENTATIVE       STATE SENATOR        1
##   unknown      party             jurisdiction jurisdictioncounty
## 1         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
## 2         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
## 3         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
## 4         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
## 5         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
## 6         REPUBLICAN LEG DISTRICT 01 - SENATE               KING
##   jurisdictiontype electionyear amount recordtype   fromdate   thrudate
## 1      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
## 2      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
## 3      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
## 4      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
## 5      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
## 6      Legislative         2012 283.25       DEBT 0012-06-01 0012-07-16
##     debtdate code         description   vendorname vendoraddress
## 1 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
## 2 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
## 3 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
## 4 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
## 5 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
## 6 0012-07-03      RE-ORDER TEE SHIRTS HICKEY GAYLE   PO BOX 2749
##     vendorcity vendorstate           candidate vendorzip
## 1 WOODINVILLE           WA         RYU CINDY S
## 2 WOODINVILLE           WA THOMAS TIMOTHY N JR
## 3 WOODINVILLE           WA         FEY JACOB C
## 4 WOODINVILLE           WA   STRACHAN STEVEN D
## 5 WOODINVILLE           WA        INSLEE JAY R
## 6 WOODINVILLE           WA   RICCELLI MARCUS M
```

Finally, we saved each dataframe as a *R* data file for use in the next steps of our EDA.

```
if (save_transformed_datatables==T)
{
  save(candidate_debt_transformed,file="candidate_debt_transformed.Rda")
  save(candidate_debt_clean,file="candidate_debt_clean.Rda")
  save(candidate_debt_bad,file="candidate_debt_bad.Rda")
}
```

Univariate Analysis of Key Variables

This is a univariate analysis of key variables. Excluded from this analysis were variables that showed names, identifiers, addresses, vendor information, and variables that are the same for all candidates (such as filertype). We excluded these, because they identify vendors and candidates, and to not represent variables relevant to debt. Also, for the time being, dates are excluded, since there is only one election cycle covered. Therefore (and unfortunately), date-related data is more of a constant than a variable.

Starting with the "Office" variable, which describes the office sought by the candidate, the summary statistics are below. Because this is a categorical variable, a bar chart or table would be appropriate visualizations. Also because of the varying sizes (State Representative has far more observations than the others combined), there might not be a meaningful relationship shown with this variable in this dataset.

```
summary(candidate_debt_transformed$office)
```

```
##          APPEALS COURT JUDGE             ATTORNEY GENERAL
##                            4                           34
##              COUNTY ASSESSOR          COUNTY COMMISSIONER
##                           19                           72
##        COUNTY COUNCIL MEMBER               COUNTY SHERIFF
##                           15                           17
##                     GOVERNOR    PUBLIC LANDS COMMISSIONER
##                           42                           27
## PUBLIC UTILITY COMMISSIONER           SECRETARY OF STATE
##                            8                           11
##               STATE AUDITOR         STATE REPRESENTATIVE
##                            7                          528
##               STATE SENATOR STATE SUPREME COURT JUSTICE
##                          118                           28
##              STATE TREASURER         SUPERIOR COURT JUDGE
##                           24                           33
```

Next, we examined the legislative district variable. Because the variable is categorical, and with many long labels, we used the summary below to show frequency.

```
summary(candidate_debt_transformed$legislativedistrict)
```

```
##             ATTORNEY GENERAL              COUNTY ASSESSOR
##                           49                            2
##          COUNTY COMMISSIONER        COUNTY COUNCIL MEMBER
##                           44                            3
##               COUNTY SHERIFF                     GOVERNOR
##                            8                          101
##    PUBLIC LANDS COMMISSIONER PUBLIC UTILITY COMMISSIONER
##                           21                            2
##               STATE AUDITOR         STATE REPRESENTATIVE
##                            4                          384
##               STATE SENATOR STATE SUPREME COURT JUSTICE
##                          305                           12
##              STATE TREASURER         SUPERIOR COURT JUDGE
##                           28                           24
```

Regarding the position variable, it appears to be a numerical code for something, but the meaning behind each code was not provided. Even if there were a relationship between position code and debt, we can't derive meaning without knowing what each of the numerical codes in position means.

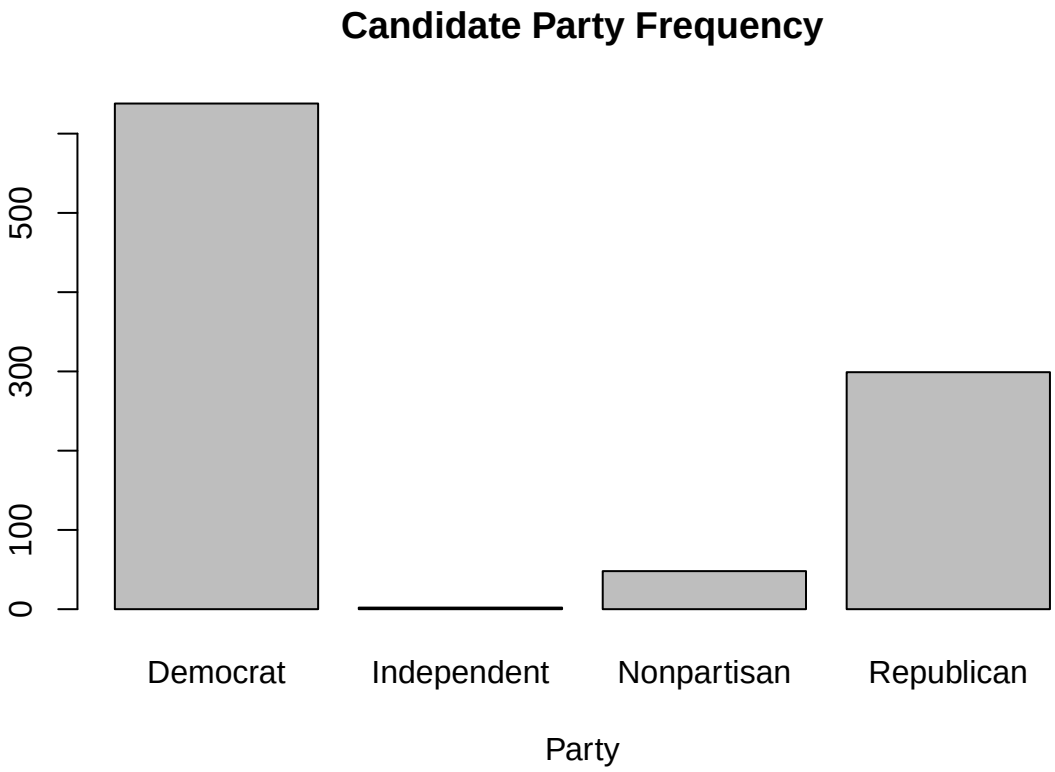
Below is a summary and bar chart representing the party variable. We used a bar chart to show frequency as

it is a categorical variable.

```
summary(candidate_debt_transformed$party)
```

```
##     DEMOCRAT  INDEPENDENT NON PARTISAN   REPUBLICAN
##          638            2           48          299
```

```
counts <- table(candidate_debt_transformed$party)
barplot(counts, main="Candidate Party Frequency",
        xlab="Party", names.arg=c("Democrat", "Independent", "Nonpartisan", "Republican"))
```



Below are summaries for the categorical variables jurisdiction and jurisdiction county. Because jurisdiction and jurisdiction county are highly likely to be correlated, we should consider this in future analysis. Specifically, a PCA or factor analysis should be considered.

```
summary(candidate_debt_transformed$jurisdiction)
```

```
## ATTORNEY GENERAL, OFFICE OF    AUDITOR, OFFICE OF STATE
##                          49                           4
##                  CLALLAM CO        GOVERNOR, OFFICE OF
##                          10                         101
##                   ISLAND CO               JEFFERSON CO
##                           6                           1
##                     KING CO      KING CO SUPERIOR COURT
##                           8                          21
##     LEG DISTRICT 01 - HOUSE    LEG DISTRICT 01 - SENATE
##                           3                         243
##     LEG DISTRICT 03 - HOUSE    LEG DISTRICT 03 - SENATE
##                           1                           7
##    LEG DISTRICT 05 - SENATE     LEG DISTRICT 08 - HOUSE
##                           3                           8
##     LEG DISTRICT 11 - HOUSE    LEG DISTRICT 11 - SENATE
```

```
##                          13                          30
##     LEG DISTRICT 14 - HOUSE     LEG DISTRICT 21 - HOUSE
##                           2                          55
##     LEG DISTRICT 22 - HOUSE     LEG DISTRICT 23 - HOUSE
##                           1                           3
##     LEG DISTRICT 24 - HOUSE     LEG DISTRICT 25 - HOUSE
##                           5                           4
##     LEG DISTRICT 26 - HOUSE     LEG DISTRICT 27 - HOUSE
##                          37                           6
##     LEG DISTRICT 28 - HOUSE     LEG DISTRICT 29 - HOUSE
##                          21                           2
##     LEG DISTRICT 32 - HOUSE     LEG DISTRICT 34 - HOUSE
##                           6                           7
##     LEG DISTRICT 36 - HOUSE     LEG DISTRICT 37 - HOUSE
##                           9                           1
##     LEG DISTRICT 40 - HOUSE    LEG DISTRICT 40 - SENATE
##                           6                          16
##     LEG DISTRICT 41 - HOUSE    LEG DISTRICT 41 - SENATE
##                           5                           5
##     LEG DISTRICT 43 - HOUSE     LEG DISTRICT 44 - HOUSE
##                          72                           1
##     LEG DISTRICT 45 - HOUSE     LEG DISTRICT 46 - HOUSE
##                          68                          17
##    LEG DISTRICT 46 - SENATE     LEG DISTRICT 47 - HOUSE
##                           1                          13
##     LEG DISTRICT 48 - HOUSE                   MASON CO
##                          18                          15
##  NATURAL RESOURCES, DEPT OF  OKANOGAN CO SUPERIOR COURT
##                          21                           1
##                   PIERCE CO    PIERCE CO SUPERIOR COURT
##                           5                           1
##                   SKAGIT CO               SUPREME COURT
##                          12                          12
##  THURSTON CO SUPERIOR COURT     THURSTON PUD 01, 02, 03
##                           1                           2
##  TREASURER, OFFICE OF STATE
##                          28
```

```
summary(candidate_debt_transformed$jurisdictioncounty)
```

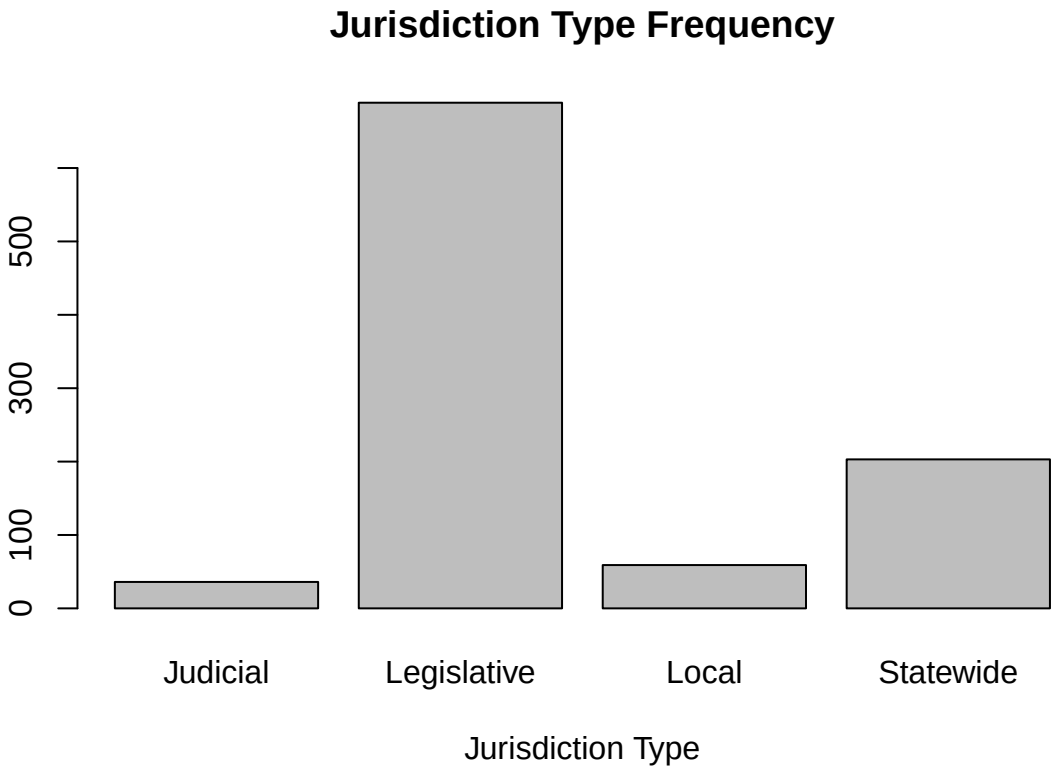
```
##              BENTON   CLALLAM    ISLAND JEFFERSON      KING    KITSAP
##       215         8        15         6         1       544         3
##     MASON  OKANOGAN    PIERCE    SKAGIT SNOHOMISH   SPOKANE  THURSTON
##        15         1        76        34        55         8         4
##    YAKIMA
##         2
```

Then, we summarized the jurisdiction type categorical variable, using a bar plot to show frequency. There seems to be some coding issues here. For instance, it isn't clear that judicial and legislative can't overlap with local and/or statewide. As an example, it is unclear how a candidate for state legislature would be coded.

```
summary(candidate_debt_transformed$jurisdictiontype)
```

```
##    Judicial Legislative       Local   Statewide
##          36         689          59         203
```

```
counts <- table(candidate_debt_transformed$jurisdictiontype)
barplot(counts, main="Jurisdiction Type Frequency",
        xlab="Jurisdiction Type")
```

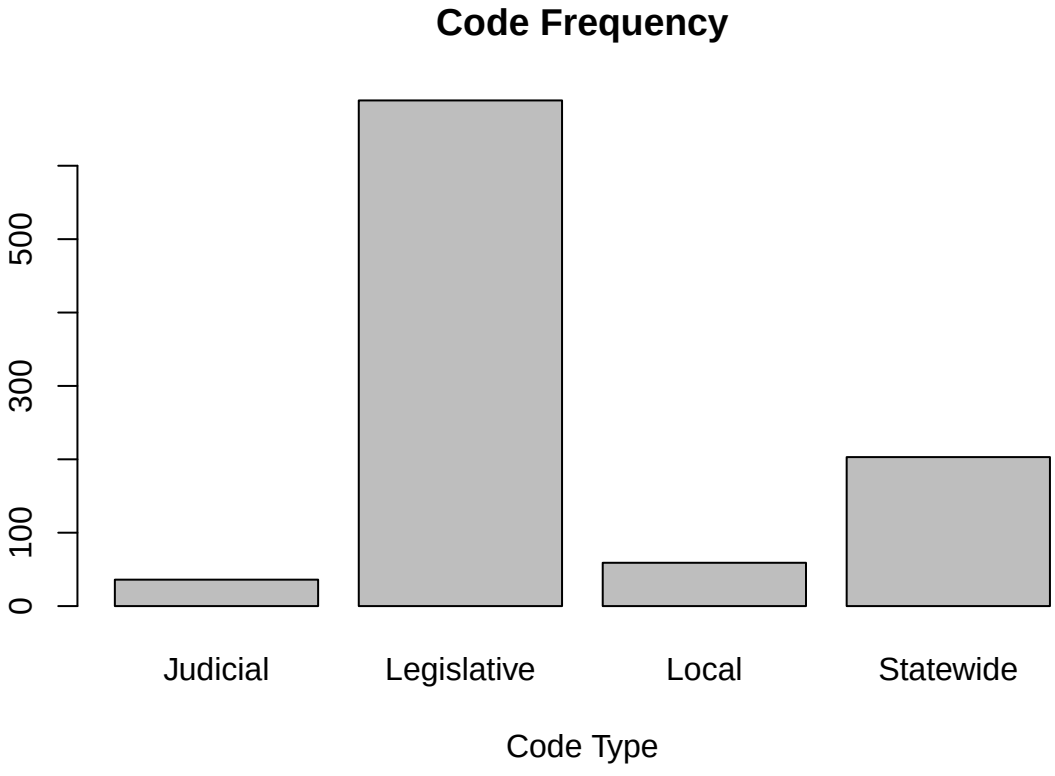


We then summarized the categorical code variable, with a bar plot below to show frequency. This shows an issue with the data, in that there are more missing values than the others combined. Considering that nearly 3/5 of the data values are missing in this variable, at best, further analysis of this would provide a very incomplete story.

```
summary(candidate_debt_transformed$code)
```

```
##                                 Fundraising    Management Services
##                     610                   5                    10
## Operation and Overhead
##                     362
```

```
barplot(counts, main="Code Frequency",
        xlab="Code Type")
```



Below is a summary of the categorical description variable. There seems to be another coding issue here. For instance, some descriptions are useful and intuitive, such as “mileage” or “office supplies”. However, descriptions like “March Treasury” don’t actually say what this is for, and further clarification is needed for analysis to be meaningful. Also, there are many categories with only 1 or 2 data points, but are similar. These could be truncated. For instance, “postage, postage/printing, and”office supplies" could be combined into “office supplies”.

```
summary(candidate_debt_transformed$description)
```

```
##    Length     Class      Mode
##       987 character character
```

```
#@Mary -- wanted to make this short so I sorted it and only listed the first 10
description = candidate_debt_transformed$description
description.freq = table(description)
description.freq.top = sort(description.freq, decreasing = T)[0:10]
cbind(description.freq.top)
```

```
##                          description.freq.top
## RE-ORDER TEE SHIRTS                       241
## CONSULTING/TRAVEL                          85
## ACCOUNTING/COMPLIANCE                      77
## NOVEMBER TREASURY                          58
##                                            39
## JUNE FUNDRAISING                           33
## REIMB. MTG EXP (MERCATO)                   31
## OCTOBER TREASURY                           21
## WIN BONUS                                  21
## SEPTEMBER TREASURY                         19
```

The final variable we examined in this univariate analysis is the only continuous relevant variable in this

dataset. Below are the summary statistics and the default histogram, incorporating all of the data.

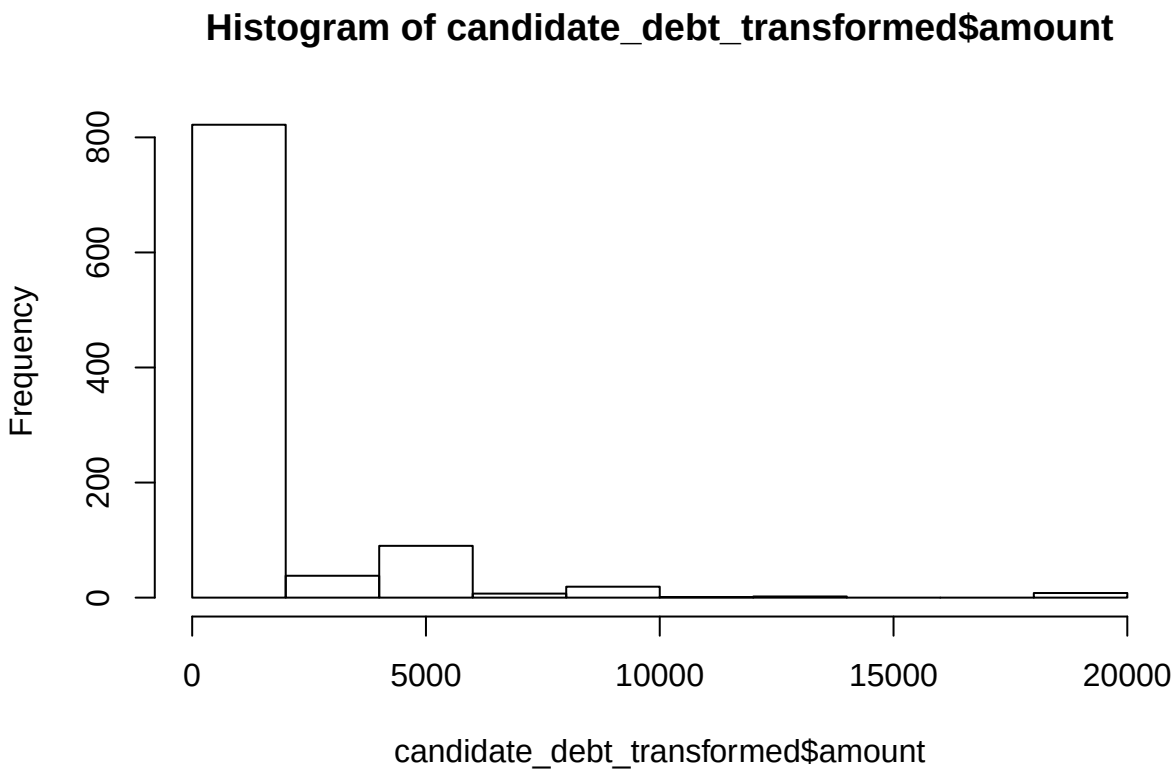
```
mean=mean(candidate_debt_transformed$amount)
median=median(candidate_debt_transformed$amount)
max=max(candidate_debt_transformed$amount)
```

Specifically, this data appears to have a striking right skew, as the mean is $1347.4247619, while the median debt is $300. This difference of over $1047, combined with a maximum of $19000 suggests that most candidates did not take on much debt, but a few went heavily into debt. This right skew is also very apparent in the histogram, but this doesn't tell us much about the majority of candidates.

```
summary(candidate_debt_transformed$amount)
```

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

```
hist(candidate_debt_transformed$amount)
```



To zoom in a bit, we created a subset of the candidate debt data and set some thresholds for our analysis. We iterated over these values until we settled at the values shown below. We have added an additional column into the data table to classify the entries based on the amount of debt recorded by the filer.

```
verylowdebt.amount <- as.integer(500)
lowdebt.amount <- as.integer(2000)
veryhighdebt.amount <- as.integer(5000)
outlierdebt.amount <- as.integer(10000)
maxdebt=as.integer(max(candidate_debt_transformed$amount))

candidate_debt_transformed <- candidate_debt_transformed %>%
  mutate(debt_category= factor(
       dplyr::case_when(amount <= verylowdebt.amount ~ "verylowdebt",
                        amount > verylowdebt.amount & amount <= lowdebt.amount ~ "lowdebt",
```

```
                        amount > lowdebt.amount & amount <= veryhighdebt.amount ~ "highdebt",
                        amount > veryhighdebt.amount  ~ "veryhighdebt"), levels=c("verylowdebt","lowdebt
```

```
(debt_category.freq <- ftable(candidate_debt_transformed$debt_category))
```

```
##  verylowdebt lowdebt highdebt veryhighdebt
##
##          646     176       44          121
```

Then we examine the records that are equal or below the lowdebt classification

```
lowdebt <- subset(candidate_debt_transformed, amount <= lowdebt.amount)
```

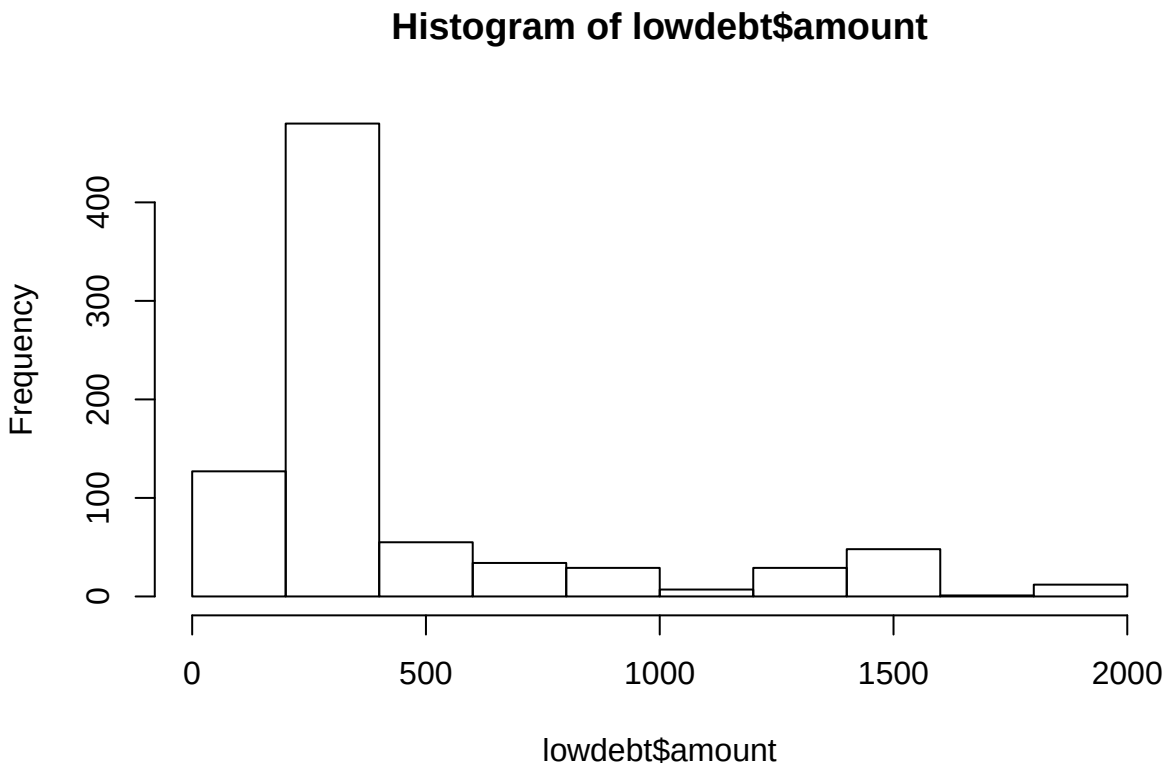
Since the cutoff from the histogram looks to be around $2000, we started there. This subset comprises 822 out of 987 observations, so looking at this in more detail makes sense.

In examining the summary statistics and histogram of this subset, the right skew is similar, but less pronounced.

```
summary(lowdebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.24  250.00  283.25  455.85  439.06 2000.00
```

```
hist(lowdebt$amount)
```



To zoom in further, we then examined the subset of candidates with debt below $500, as this appears to be the next cutoff.

```
verylowdebt <- subset(candidate_debt_transformed, amount <= verylowdebt.amount)
```

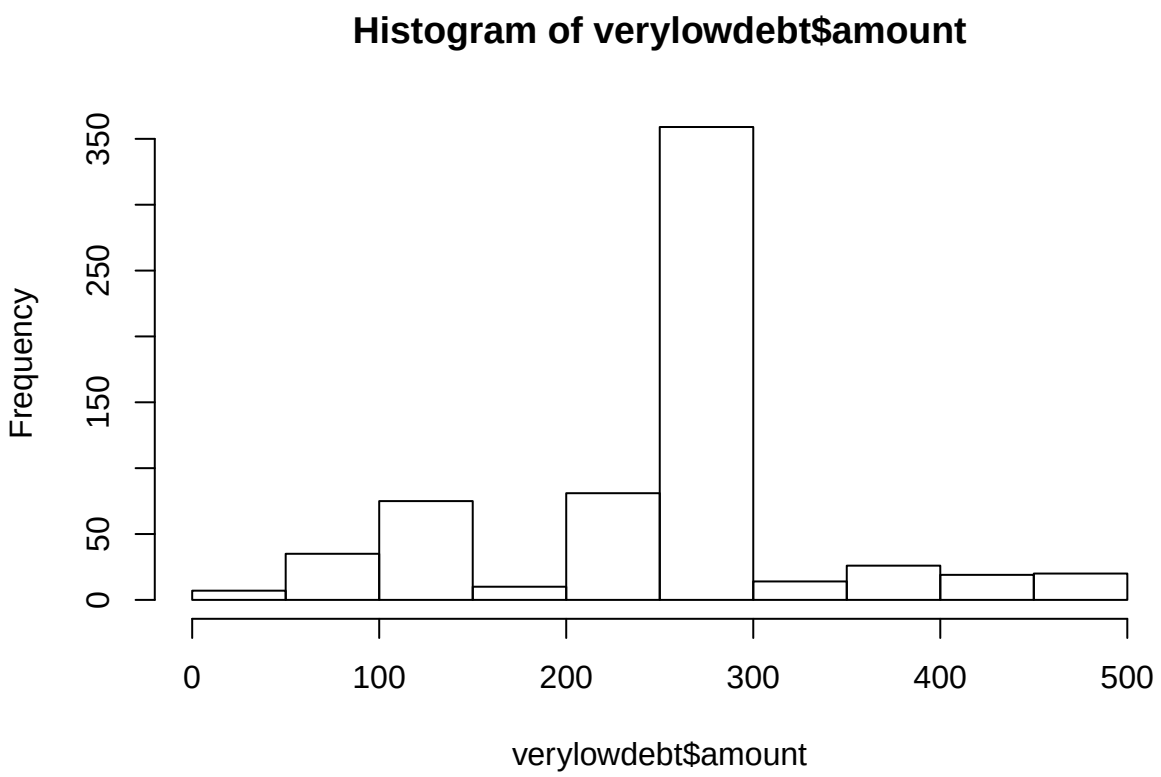
This subset comprises 646 out of a total of 987 observations, roughly 2/3 of the total. Therefore, a deeper look is warranted.

The summary statistics show a very slight left skew -0.2212763, with only $22 difference between median and mean. Looking at the histogram, this does not look normally distributed, as it appears to have a very high kurtosis 3.9058341 (it looks pointy). However, the data for roughly 2/3 of the candidates look quite different than the data for all candidates. This is important to keep in mind for post-EDA research.

```
summary(verylowdebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.24  225.00  283.25  261.38  300.00  500.00
```

```
hist(verylowdebt$amount)
```

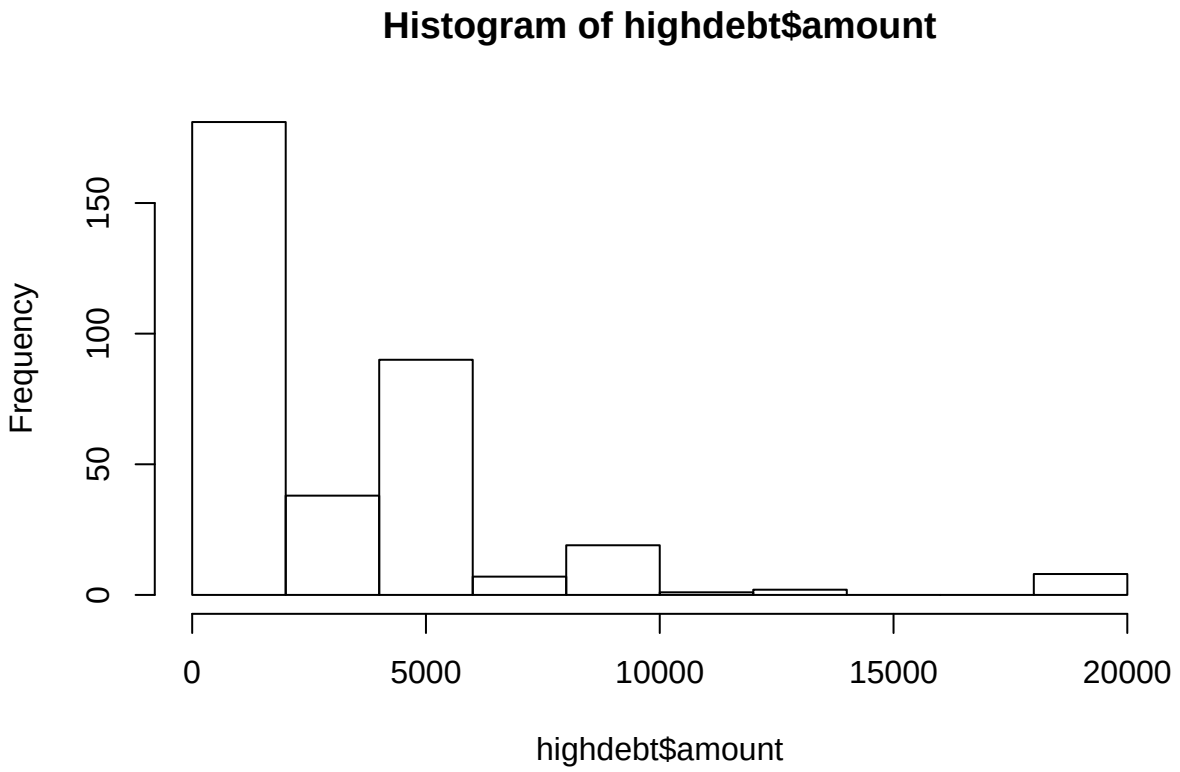


Since 1/3 of the candidates owe between $500 and $19000, they are also worth looking into. For the roughly 1/3 of candidates who owed more than $500, there appears to be a large right skew in the distribution. From looking at the histogram, the skew looks less pronounced than the original histogram, but there is a sharp dropoff at around $5000.

```
highdebt <- subset(candidate_debt_transformed, amount >= verylowdebt.amount)
summary(highdebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500    1094    1982    3363    5070   19000
```

```
hist(highdebt$amount)
```

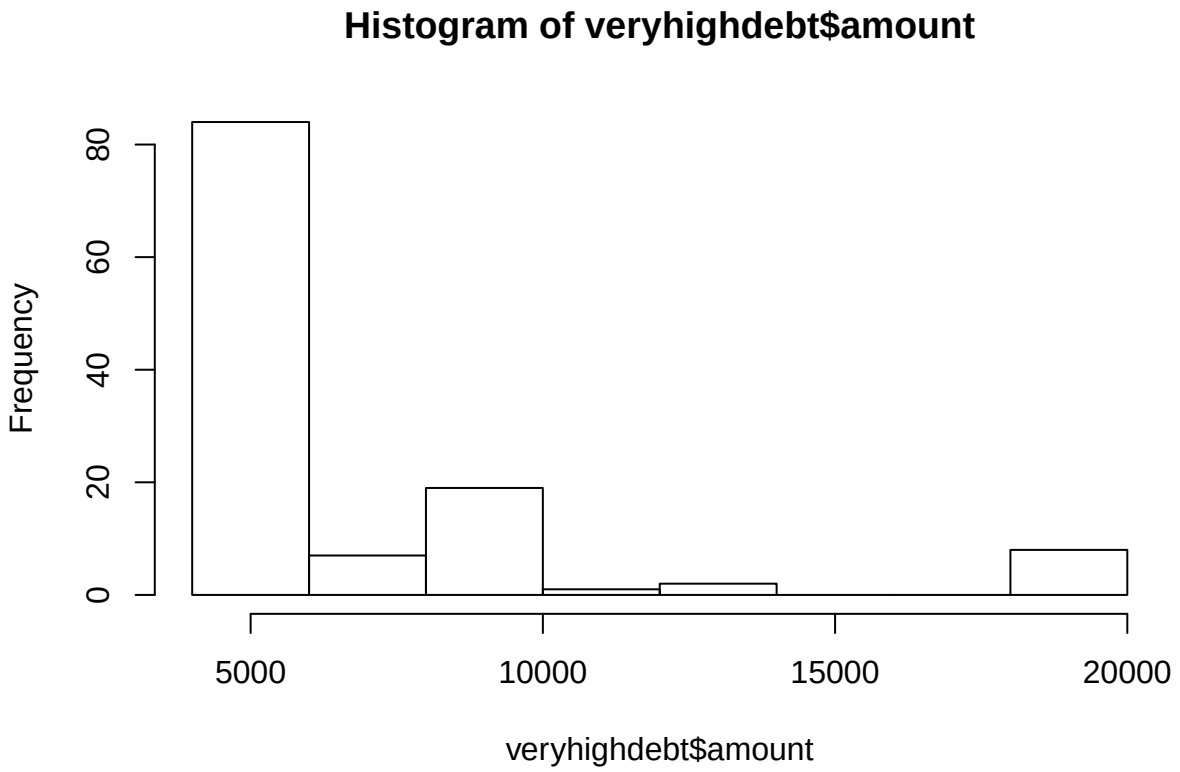


While those holding more debt than $5000 might be outliers, we created a subset and to examine the data of candidates with very high debt. This subset is 346 out of 987 total candidates. While it is a minority, it's a stretch to call roughly 1/9 of the data outliers. Also, the summary statistics and the histogram show a sharp right skew, and a cutoff at around $10000.

```
veryhighdebt <- subset(candidate_debt_transformed, amount >= veryhighdebt.amount)
summary(veryhighdebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5070    5070    5070    6767    7219   19000
```

```
hist(veryhighdebt$amount)
```

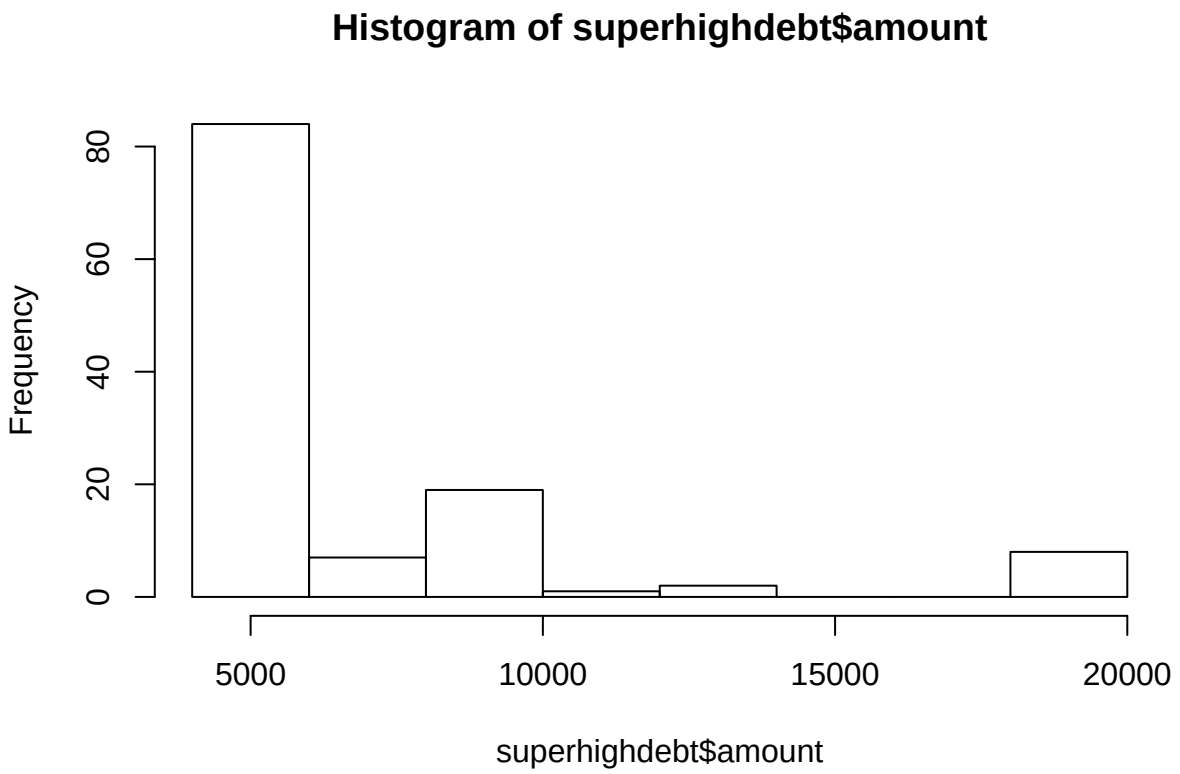


Zooming in even farther, looking into candidates with over $10000 in debt, we only found 121 out of 987 observations. Therefore, it is safe to treat these as outliers. The summary statistics and histogram shows a left skew, but several with $19000 in debt.

```
superhighdebt <- subset(candidate_debt_transformed, amount >= veryhighdebt.amount)
summary(superhighdebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5070    5070    5070    6767    7219   19000
```

```
hist(superhighdebt$amount)
```

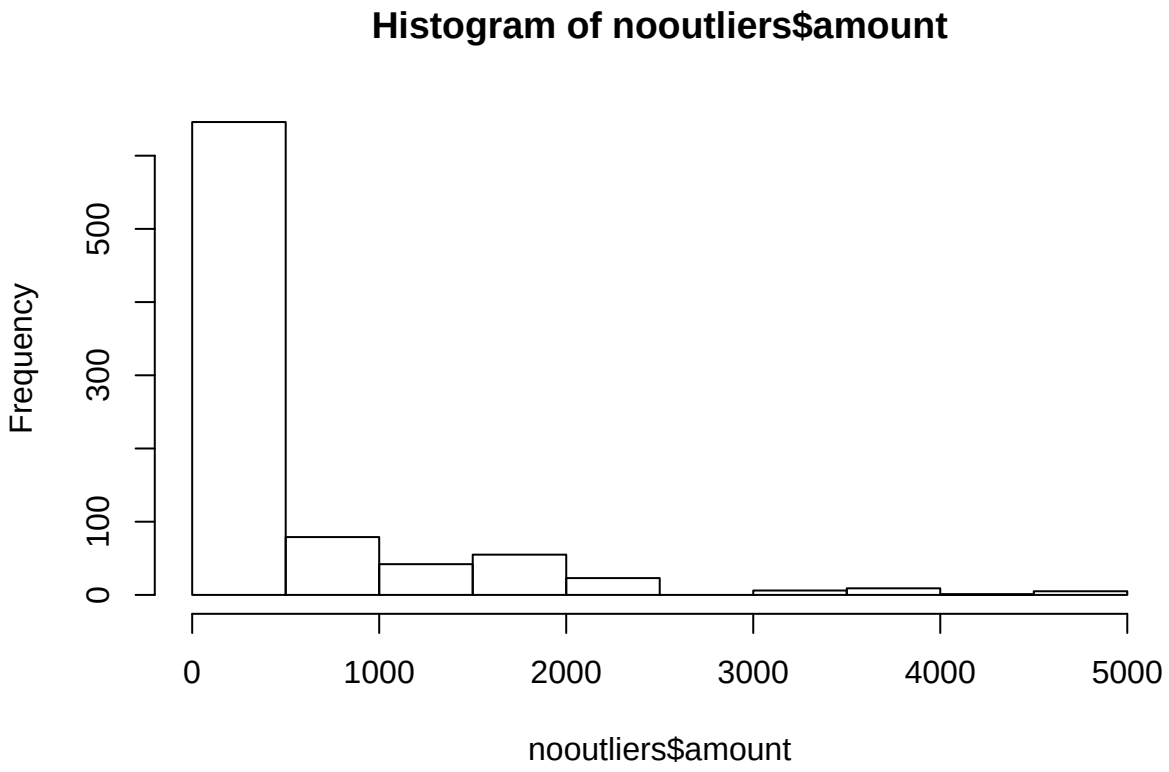


We then examined the data, excluding outliers. Taking out the outliers reduced some, but not all of the skew, and the histogram doesn't look much different than the original.

```
nooutliers <- subset(candidate_debt_transformed, amount <= veryhighdebt.amount)
summary(nooutliers$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.24  283.25  283.25  590.20  560.29 4609.00
```

```
hist(nooutliers$amount)
```

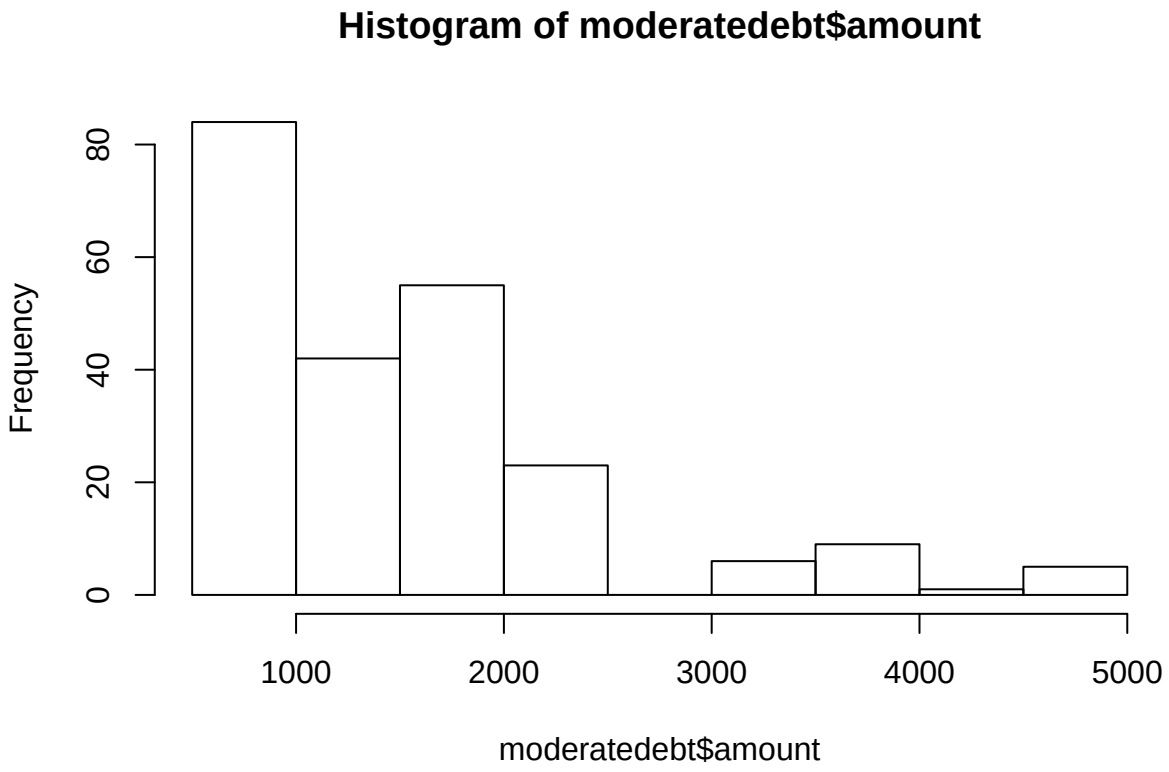


Our final look into the amount data involved candidates with debt between $500 and $10000, excluding the majority with low debt and the outliers. Below we can see something interesting. This moderate debt distribution looks quite different than that with low debt or high debt. For instance, it is clearly bimodal with a sharp right skew, with cutoff points at roughly $2K and $6K. For post-EDA analysis, we should see what might explain the cutoff points of: $500, $`lowdebt.amount`, $5000, and $10000.

```
moderatedebt <- subset(candidate_debt_transformed, amount >= verylowdebt.amount & amount <= veryhighdeb
summary(moderatedebt$amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   500.0   897.8  1220.3  1532.3  1700.0  4609.0
```

```
hist(moderatedebt$amount)
```

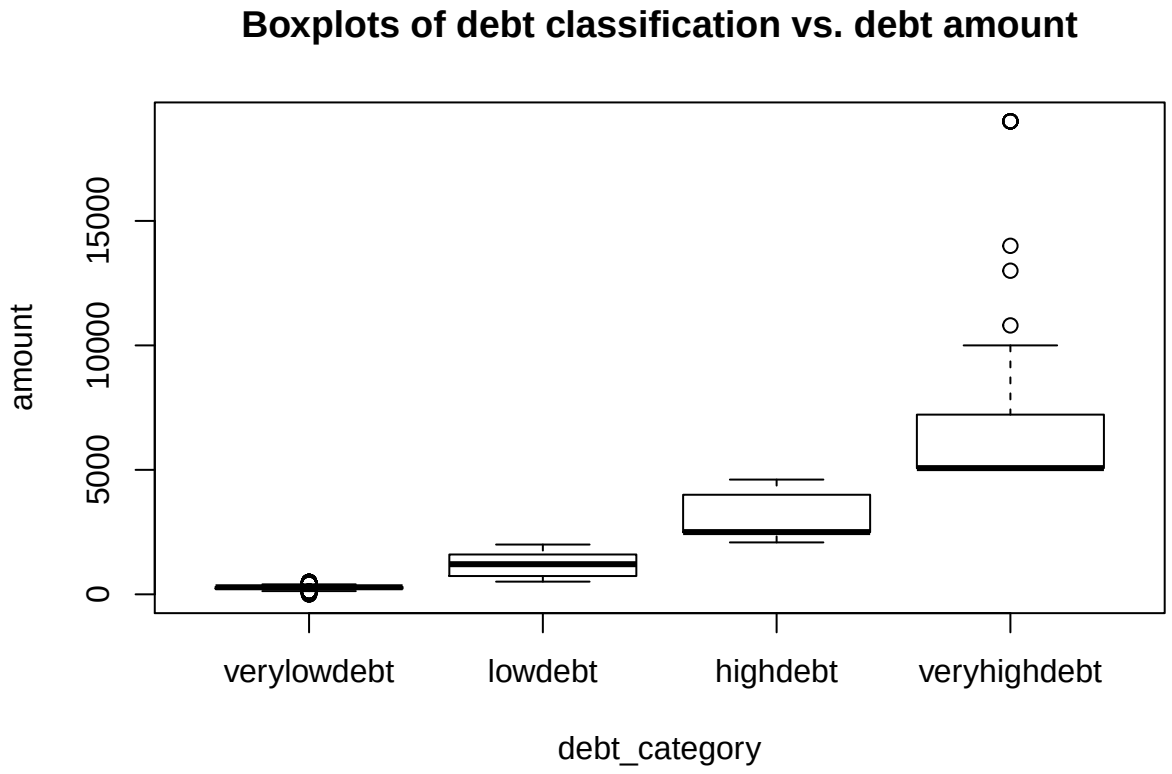


Analysis of Key Relationships

In this section, we explored how our outcome variable is related to the other variables in this dataset and presented visualizations to show the nature of each bivariate relationship.

Below is an initial box plot that includes outliers for the debt classification versus debt amount. As we can see, the outliers are primarily those with very high debt.

```
Boxplot(amount ~ debt_category,data=candidate_debt_transformed,id.method="n",id.n=3,main="Boxplots of d
```

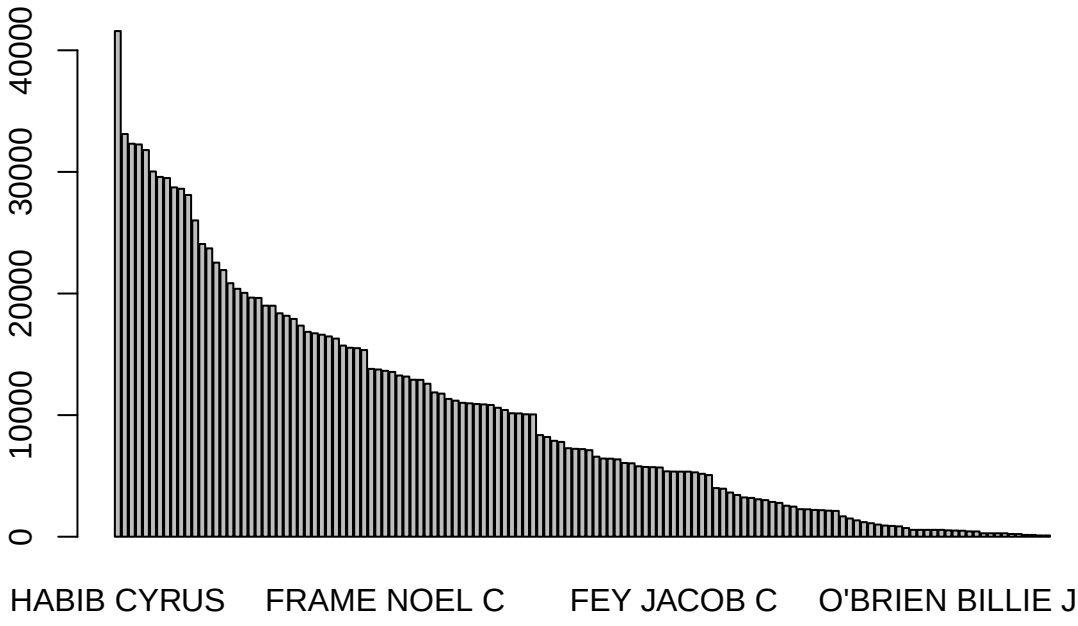



```
#changing filername to 'candidate' because that's what it is and

candidate_debt_transformed$amount = as.numeric(candidate_debt_transformed$amount)
```

First, we looked at the relationship of debt to candidate. The bar chart shows that Habib Cyrus is a slight outlier had the most debt, but nothing else seems surprising.

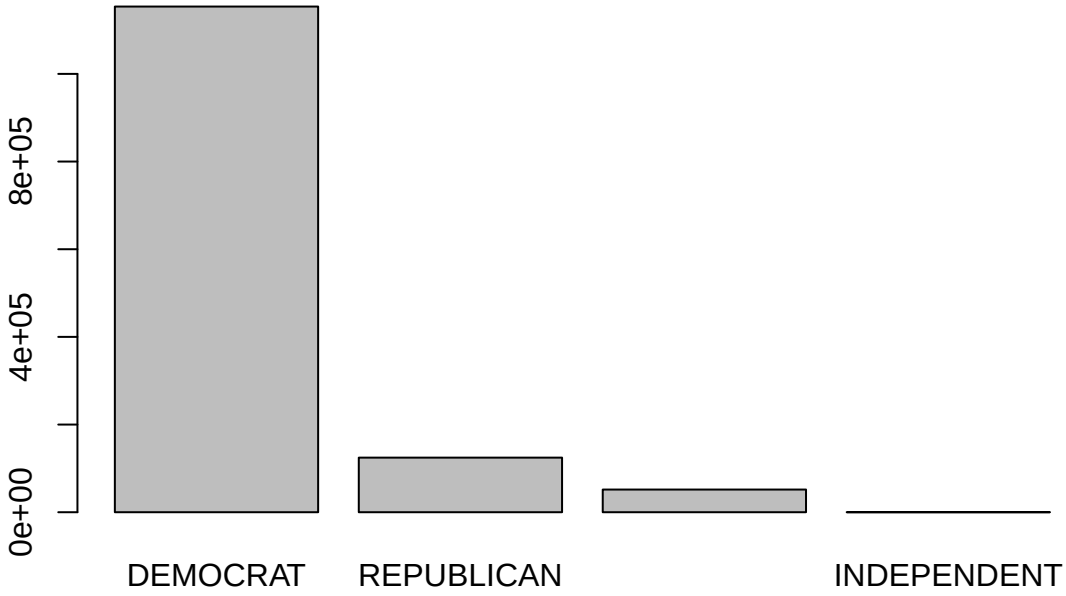
```
#Let's take a look at total debt by candidate
cand_sum = sort(by(candidate_debt_transformed$amount, candidate_debt_transformed$candidate, sum), decrea
barplot(cand_sum)
```



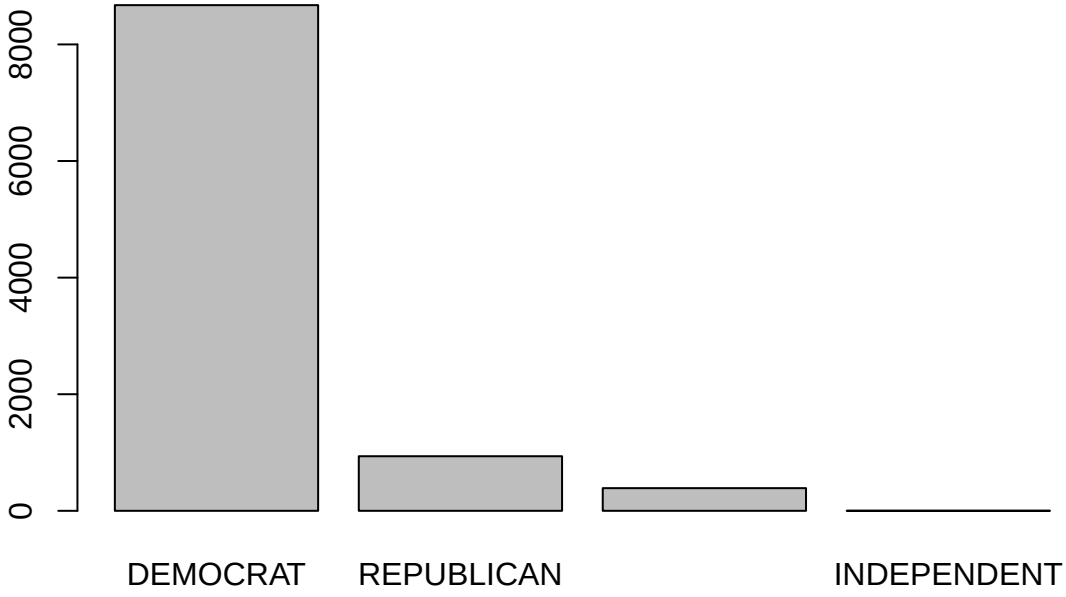
Next, we

checked the total debt amount by party. While the data shows democratic candidates far more debt than any other party, it seems likely that this is a result of there the distribution of party affiliation discussed in the uni-variate analysis. To check this, we divide the total party debt by the number of candidate for each party. This will show us the average debt per candidate.

```
#Let's also check a look at total debt by party
party_sum = sort(by(candidate_debt_transformed$amount, candidate_debt_transformed$party, sum), decreasi
barplot(party_sum)
```




```
#Let's also check a look at debt by party divide by number of unique candidats
party_avg = sort(by((candidate_debt_transformed$amount)/length(unique(candidate_debt_transformed$candida
barplot(party_avg)
```




```
party_avg
```

```
## candidate_debt_transformed$party
##     DEMOCRAT   REPUBLICAN NON PARTISAN  INDEPENDENT
## 8673.6901504  936.1541353  388.6922556    0.7735338
```

From the bar chart, it's clear that democratic candidates have far more debt per candidate, than the next party (Republicans). While the average debt is interesting, we also looked at the distribution of debt by party.

```
mytable = table(candidate_debt_transformed$party, candidate_debt_transformed$debt_category)
print(mytable)
```

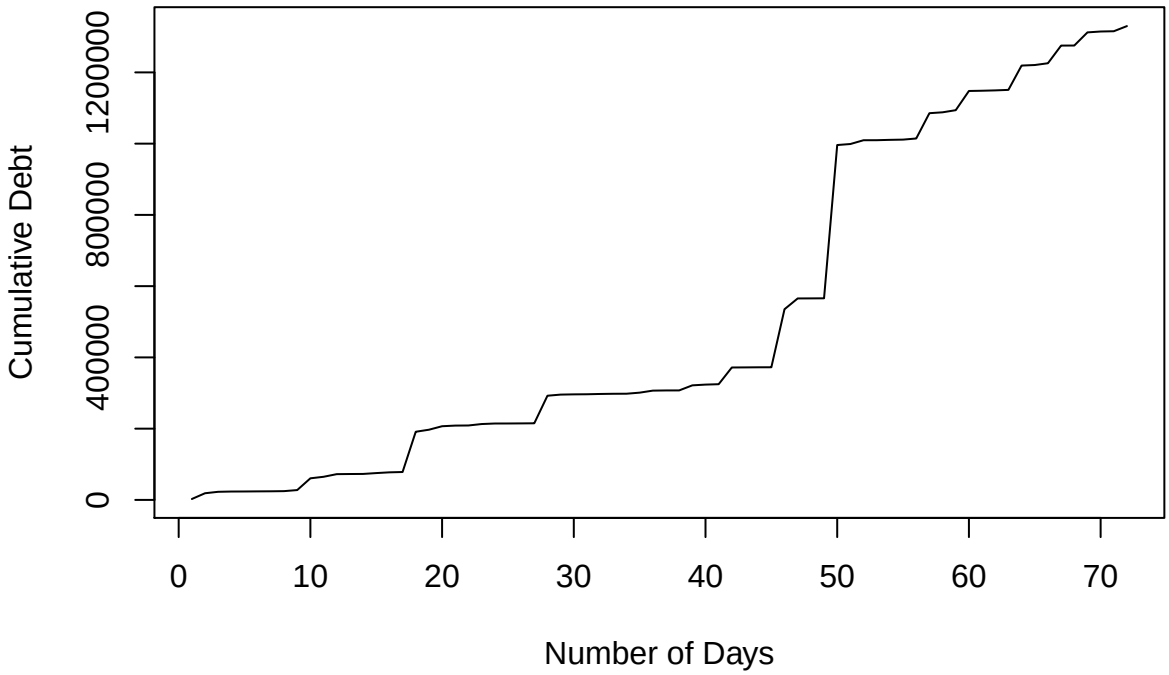
```
##
##                verylowdebt lowdebt highdebt veryhighdebt
##   DEMOCRAT             339     155       24          120
##   INDEPENDENT            2       0        0            0
##   NON PARTISAN          26      10       12            0
##   REPUBLICAN           279      11        8            1
```

From the table we see that the distribution of democrats by category is somewhat skewed, with records tending to have low/very low debt or very high debt. In contrast, Republicans only have a significant number of records in the very low category.

Next, we examined debt over time. As we can see below, debt increases over time, as expected.

```
candidate_debt_transformed$my_date = as.Date(candidate_debt_transformed$debtdate, format='%m/%d/%Y')
my_var = by(candidate_debt_transformed$amount,candidate_debt_transformed$my_date, sum)

plot(cumsum(my_var), type="l",
     xlab="Number of Days",ylab="Cumulative Debt")
```



Next, we looked at top categories that the debt funded. Consulting not only had the most total debt attributed, but also ranked highest in amount per record.

```
#desc_amount = sort(by(candidate_debt_transformed$amount, candidate_debt_transformed$description, sum),

#print(desc_amount[0:5])

# We can also check what descriptions have the highest avg amount
#avg_desc_amount = sort(by(candidate_debt_transformed$amount, candidate_debt_transformed$description, m
```

```
#print(avg_desc_amount[0:5])
```

Since so much money was spent on consulting we decided to examine the vendor states for consulting. After filtering to only debt with a description of Consulting, we summed the amount and found that 152,000 was spent on DC consultants, and only 28,250 on local Washington State consultants. The is in contrast to the overall totals, where ~690,000 was spent in Washington State to ~582,000 being spent in DC.

```
couns = subset(candidate_debt_transformed, description == "CONSULTING")
c_states = sort(by(couns$amount, couns$vendorstate, sum), decreasing = T)
print(c_states)
```

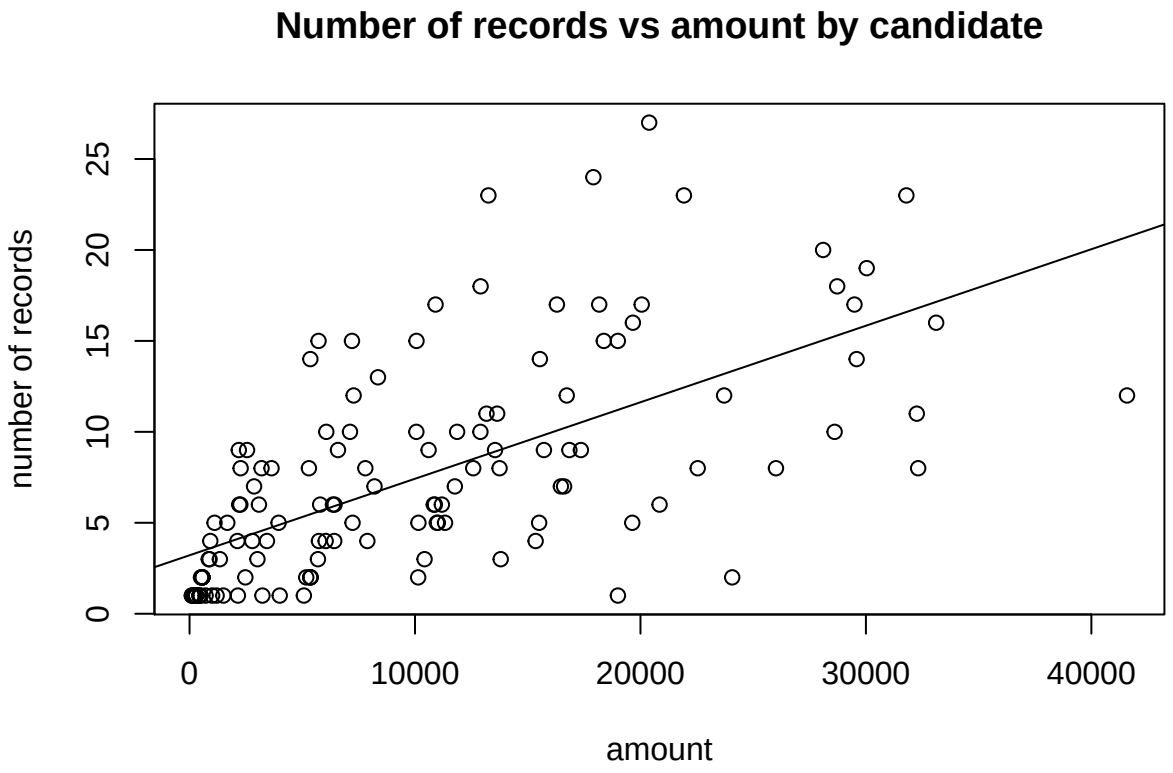
```
## couns$vendorstate
##     DC     WA
## 152000  28250
```

```
states = sort(by(candidate_debt_transformed$amount, candidate_debt_transformed$vendorstate, sum), decrea
print(states)
```

```
## candidate_debt_transformed$vendorstate
##        WA        DC        CA                TX
## 690493.39 582816.76  38880.15  17323.16    394.78
```

Finally, we examined the relationship between the number of records per candidate and the total amount of debt.

```
x_var = by(candidate_debt_transformed$amount, candidate_debt_transformed$candidate, sum)
y_var = by(candidate_debt_transformed$reportnumber, candidate_debt_transformed$candidate, length)
plot(x_var, y_var,
     xlab = "amount", ylab = "number of records",
     main = "Number of records vs amount by candidate")
abline(lm(y_var ~ x_var))
```



```r
summary((lm(y_var ~ x_var)))
```

```
##
## Call:
## lm(formula = y_var ~ x_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.340  -2.514  -1.256   2.554  15.210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.215e+00  5.850e-01   5.495 1.96e-07 ***
## x_var       4.207e-04  4.289e-05   9.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.588 on 131 degrees of freedom
## Multiple R-squared:  0.4234, Adjusted R-squared:  0.4189
## F-statistic: 96.17 on 1 and 131 DF,  p-value: < 2.2e-16
```

While there clearly is a relationship, the adjusted $R^2 = .42$ means that less than half of the variance is explained by the number of records. This means that there are clearly a lot of other factors that impacted how much debt candidates took on. With our current data set one factor we can test is party affiliation. To test this, we ran the same regression with only democrats and only Republicans. We found that for Republicans the adjusted R^2 dropped to .16. For Democrats, the adjusted R^2 at .31 was higher than Republicans but lower than our original calculation of .42.

```r
candidate_debt_transformed_r = subset(candidate_debt_transformed, party == "REPUBLICAN")
candidate_debt_transformed_d = subset(candidate_debt_transformed, party == "DEMOCRAT")


x_var_r = by(candidate_debt_transformed_r$amount, candidate_debt_transformed_r$candidate, sum)
y_var_r = by(candidate_debt_transformed_r$reportnumber, candidate_debt_transformed_r$candidate, length)
summary((lm(y_var_r ~ x_var_r)))
```

```
##
## Call:
## lm(formula = y_var_r ~ x_var_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6487 -1.3670 -0.6172  1.0337  5.1609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2680775  0.2463584   9.206 7.54e-15 ***
## x_var_r     0.0006163  0.0001389   4.435 2.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.701 on 96 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1614
## F-statistic: 19.67 on 1 and 96 DF,  p-value: 2.45e-05
```

```r
x_var_d = by(candidate_debt_transformed_d$amount, candidate_debt_transformed_d$candidate, sum)
y_var_d = by(candidate_debt_transformed_d$reportnumber, candidate_debt_transformed_d$candidate, length)
summary((lm(y_var_d ~ x_var_d)))
```

```
##
## Call:
## lm(formula = y_var_d ~ x_var_d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.865 -2.463 -1.262  1.994 13.489
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.462e+00  6.082e-01   4.048 9.70e-05 ***
## x_var_d     3.162e-04  4.405e-05   7.177 9.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.217 on 109 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.3209, Adjusted R-squared:  0.3147
## F-statistic: 51.51 on 1 and 109 DF,  p-value: 9.13e-11
```

Analysis of Secondary Effects

In this data set, the most interesting relationship is between party and debt. However, there is reason to hypothesize that this is actually a secondary effect. Specifically, we are lacking any data on whether or not each candidate was an incumbent or a challenger. It is reasonable to hypothesize that challengers would spend more and be less able to raise funds, so therefore hold more debt. Holding this constant, party may or may not matter, especially in a state that is dominated by one party over the other.

Conclusion

This data set shows some interesting patterns. For instance, the amount of debt that candidates hold appears to have a relationship to their political party. Especially since this is a bit surprising, more analysis would be helpful.... To answer these questions analysed the average, and distribution of debt bins by party. We found that democrats are have much more debt on average, and are far more likely to be in the very high debt category than republicans. Most candidates owe less than $500, however, this amount goes up to $19000, with cutoff points at $500, $2000, $5000, and $10000. Further analysis would be useful in examining any possible underlying dynamics that might explain this pattern.

In relationship to timing we found that on X date debt seemed to spike. Because this date is close to X in the election cycle, we speculated that this might have been the cause.

We also found that the consulting was the category candidates spent most on, and than Consulting vendors from DC were paid more than local consultants. This was out of the normal trend, because in general, more money was spent with vendors in Washington state than anywhere else.

Finally we looked at the relationship between the number of records (or instances of debt being taken out) and the total amount taken out. While we found a statistically significant relationship, the R^2 below 50%

signals that other factors are might be driving the total amount. To examine these other factors we looked at the relationship within the major parties, but this only result in lower R^2 values.

Unfortunately, there was more we couldn't say than could, due to missing data, the incomplete data set, and coding issues. Specifically, if the data set included multiple years, we could see candidate debt over time. It would be helpful, also, to see which candidates hold/roll over debt, and who tends to go into debt repeatedly. It would also be extremely helpful to see fundraising numbers, coding issues clarified (especially in regards to debt description), and if the candidate was an incumbent or challenger. Finally, election outcome data would be helpful to see if there is any correlation between debt and success in an election.