Exploratory Analysis of Candidate Debt Data

Kyle Chuang, Tennison Yu, Alex Dauenhauer

May 27, 2018

Abstract

The objective of this exploratory data analysis is to better understand how campaign characteristics are related to candidate's debt in the state of Washington.

The data set provided had several issues including multiple NAs and misaligned columns. Column misalignments were able to be corrected. Majority of the NAs in column were results of local or statewide office where certain features are not applicable. The data was replaced with Local or Statewide to aid in our exploration of our key variables and relationships.

In addition to the amount of debt, the other key variables for this dataset that we wanted to explore were grouped by political variables ('office', 'debtdate', 'party', and 'position'), geographic variables ('jurisdiction' and 'legislativecounty') and financial description variables ('description', 'vendorname', 'vendorcity', and 'vendorstate').

Another primary issue affecting the exploratory data analysis (EDA) was the inability to summarize the data by candidate (filerID feature) due to candidates switching parties, positions, legislative district, etc. Nevertheless, there were many interesting observations that suggests that Democrats strongly leverage their available resources and casts a wide, perhaps controversial, net to win all the roles available. Conversely, Republicans seem to have a more focused strategy in terms of spending behavior of the candidates that were looked at. Some observations of note include Republican's predilection for T-shirt spending and the 1 Democratic candidate providing financing services for almost all candidates in the dataset.

Introduction

The objective of this analysis is to better understand how campaign characteristics are related to candidate's debt in the state of Washington. To help answer this question, we have explored a set of candidate debt data from the 2012 election year in the state of Washington. The first step is to take a look at the data which is housed in the "CandidateDebt.csv" file.

```
path = choose.dir()
setwd(path)
df = read.csv('CandidateDebt.csv')
summary(df)
```

From this we can see that there are quite a few issues with this data set that we need to clean up. The first thing we noticed was that there are no true NA values, they are inputted as '#N/A' and empty strings. We can replace the strings with true NA values so we can work around them later. We also want to drop any '#N/A's from our dataframe and perform sanity checks to make sure we didn't miss anything.

```
any(is.na(df))
any(df == "#N/A")
any(df == "")
df[df == '#N/A'] = NA
df[df == ''] = NA
df = droplevels(df)
```

The next thing we noticed was that some of the column headers did not appear to match the data in their column. First, the 'position' and 'legislativedistrict' columns appear to be swapped. To verify this, we extracted the jurisdictioncounty number and checked it against the legislativecounty. All entries matched. The code to do this is shown below.

```
legislativecounty = str_extract_all(df$jurisdictioncounty, "[0-9]+")
for (i in seq(1, length(legislativecounty))) {
  if (identical(legislativecounty[i], character(0))){
    legislativecounty[i] = NA;
  }
  else {
    # these 2 columns had 2+ districts so just pulled it to the 1st one (1)
    if (i == 852 || i == 936) {
      legislativecounty[i] = 1;
    }
    legislativecounty[i] = as.numeric(legislativecounty[i]);
  }
}
df$position = droplevels(df$position)
legislativecounty = unlist(legislativecounty, use.names=F)
all(legislativecounty == as.numeric(levels(df$position)[df$position]), na.rm=T)
rm(legislativecounty)
```

There was an additional column of empty data inserted at the 'party' variable and all variables headers after that point were shifted by one. We named the unknown column 'WHAT_IS_THIS' and it primarily acts as a placeholder which we will delete later. To fix this we realigned the headers with the correct data and dropped 'vendorzip' since there was no data associated with it. Additionally, we typecast the date variables as dates and the 'amount' variable as numeric. Then we want to remove NA's from 'amount' and perform quick sanity checks on the data.

```
names(df) = c(names(df)[1:9], names(df)[11], names(df)[10],
              'WHAT_IS_THIS', names(df)[13:length(names(df))-1])

df$amount = as.numeric(levels(df$amount)[df$amount])
df = df[!is.na(df$amount),]
df$fromdate = as.Date(levels(df$fromdate)[df$fromdate], '%m/%d/%y')
df$thrudate = as.Date(levels(df$thrudate)[df$thrudate], '%m/%d/%y')
df$debtdate = as.Date(levels(df$debtdate)[df$debtdate], '%m/%d/%y')

df$WHAT_IS_THIS = NULL
str(df[, grep('date', names(df))])
all(df$fromdate < df$thrudate, na.rm=T)
all(df$amount > 0, na.rm=T)

str(df)
```

We also noticed missing data in 'legislativedistrict' coincided with missing 'jurisdictioncounty'. Looking at the data closely, missing data in 'legislativedistrict' were caused by 'jurdisctiontype' having a value of statewide, local or judicial. We backfilled 'legislativedistrict' missing data with statewide, local or judicial to match 'jurisdictiontype'. With 'jurisdictioncounty', the situation was similar except it had missing data due to 'jurisdictiontype' being statewide or judicial. We performed a similar backfill of data with the feature and included sanity checks

```
unique(df[is.na(df$legislativedistrict),"jurisdictiontype"])
```

```
## [1] Statewide Local     Judicial
```

```
## Levels: Judicial Legislative Local Statewide
```

```
levels(df$legislativedistrict) =
  c(levels(df$legislativedistrict), 'Statewide', 'Local', 'Judicial')
df[is.na(df$legislativedistrict),"legislativedistrict"] =
  df[is.na(df$legislativedistrict),"jurisdictiontype"]
```

```
unique(df[is.na(df$jurisdictioncounty),"jurisdictiontype"])
```

```
## [1] Statewide Judicial
## Levels: Judicial Legislative Local Statewide
```

```
levels(df$jurisdictioncounty) =
  c(levels(df$jurisdictioncounty), 'Statewide','Judicial')
df[is.na(df$jurisdictioncounty), "jurisdictioncounty"] =
  df[is.na(df$jurisdictioncounty), "jurisdictiontype"]
```

We then looked at the financial description variables. There were also missing values describing how the debt was spent. Upon examination of the 'vendorname' in relation to the transaction description, it appears that the NA values for description correspond to when a credit card was used for the transaction. We filled the missing data for description with a custom factor - "Credit Card" - and performed sanity checks post replacement.

```
(credit_card_names = unique(df[is.na(df$description),]$vendorname))
credit_card_names = droplevels(credit_card_names)
df[!is.na(df$description) & (df$vendorname %in% credit_card_names), ]
levels(df$description) = c(levels(df$description), "Credit Card")
df[is.na(df$description), "description"] = "Credit Card"
```

The dataframe outputted after the last step is considered our final working product. There are other NA values associated with middle initial, vendor address, vendor state, and code. We decided it would not be appropriate to fill these in as there is no real pattern to discern what they could be filled in as.

Univariate Analysis of Key Variables

Once we cleaned up some of the variables, we wanted to look at trends in the main dependent variable, the amount of debt filed. In addition to the amount of debt, the other key variables for this dataset that we wanted to explore were grouped by political variables ('office', 'debtdate', 'party', and 'position'), geographic variables ('jurisdiction' and 'legislativecounty') and financial description variables ('description', 'vendorname', 'vendorcity', and 'vendorstate').

To start off, we looked at the mean, median, minimums, maximums and quartiles of all the data fields.

```
summary(df)
```

```
##   reportnumber       origin          filerid        filertype
##  Min.   :100346104   B.3:987   GOLDP  840: 27   Candidate:987
##  1st Qu.:100446276             MCINJ  115: 24
##  Median :100471547             BROWL  204: 23
##  Mean   :100466089             CHOPF  103: 23
##  3rd Qu.:100494036             FERGR  115: 23
##  Max.   :100599472             INSLJ  110: 20
##                                (Other)   :847
##              filername     firstname   middleinitial     lastname
##  GOLDMARK PETER J : 27   JOHN   : 42   J      :108   GOLDMARK: 27
##  MCINTIRE JAMES L : 24   ROBERT : 39   S      :106   MCINTIRE: 24
```

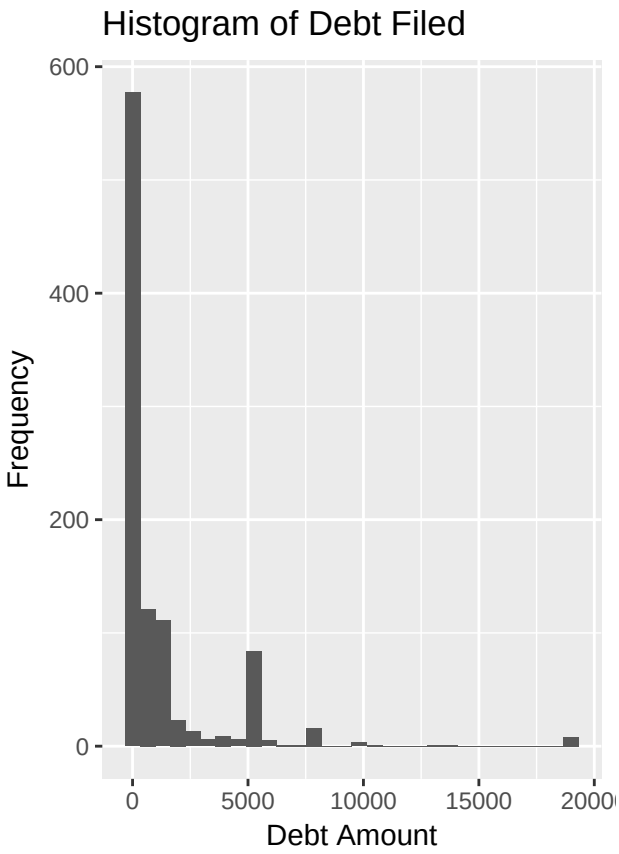
```
##  BROWN LISA J     : 23   JAMES  : 33   M      : 95   BROWN   : 23
##  CHOPP FRANK V    : 23   KEVIN  : 28   R      : 94   CHOPP   : 23
##  FERGUSON ROBERT W: 23   TIMOTHY: 28   L      : 92   FERGUSON: 23
##  INSLEE JAY R     : 20   PETER  : 27   (Other):427   FARRELL : 20
##  (Other)          :847   (Other):790   NA's   : 65   (Other) :847
##                   office                    position   legislativedistrict
##  STATE REPRESENTATIVE:528   STATE REPRESENTATIVE:384   1        :246
##  STATE SENATOR       :118   STATE SENATOR       :305   Statewide:203
##  COUNTY COMMISSIONER : 72   GOVERNOR            :101   43       : 72
##  GOVERNOR            : 42   ATTORNEY GENERAL    : 49   45       : 68
##  ATTORNEY GENERAL    : 34   COUNTY COMMISSIONER : 44   Local    : 59
##  SUPERIOR COURT JUDGE: 33   STATE TREASURER     : 28   21       : 55
##  (Other)             :160   (Other)             : 76   (Other)  :284
##           party                          jurisdiction jurisdictioncounty
##  DEMOCRAT    :638   LEG DISTRICT 01 - SENATE   :243   KING     :544
##  INDEPENDENT :  2   GOVERNOR, OFFICE OF        :101   Statewide:203
##  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)  : 60
##     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.   :2009-10-01   Min.   :2009-10-31   Min.   :2008-10-29
##  1st Qu.:2011-10-01   1st Qu.:2011-10-31   1st Qu.:2011-07-03
##  Median :2012-02-01   Median :2012-02-29   Median :2012-02-29
##  Mean   :2011-12-19   Mean   :2012-01-20   Mean   :2011-12-13
##  3rd Qu.:2012-06-01   3rd Qu.:2012-07-16   3rd Qu.:2012-07-03
##  Max.   :2012-08-01   Max.   :2012-08-31   Max.   :2012-08-31
##
##                       code                    description
##  Fundraising           :  5   RE-ORDER TEE SHIRTS  :241
##  Management Services   : 10   CONSULTING/TRAVEL    : 85
##  Operation and Overhead:362   ACCOUNTING/COMPLIANCE: 77
##  NA's                  :610   NOVEMBER TREASURY    : 58
##                               Credit Card          : 39
##                               JUNE FUNDRAISING     : 33
##                               (Other)              :454
##                      vendorname                          vendoraddress
##  HICKEY GAYLE               :241   PO BOX 2749                   :241
##  ARGO STRATEGIES            :221   PO BOX 9100                   :211
##  HIRSCHBERG STRATEGIES INC. : 84   1010 VERMONT AVE NW #814      : 84
##  PROJECT ACCOUNTING SERVICES: 79   603 STEWART ST STE 819        : 55
##  CHOPP FRANK V              : 44   1414 DEXTER AVENUE N SUITE 210: 40
##  RUDERMAN CONSULTING        : 38   (Other)                       :332
##  (Other)                    :280   NA's                          : 24
##         vendorcity  vendorstate
```

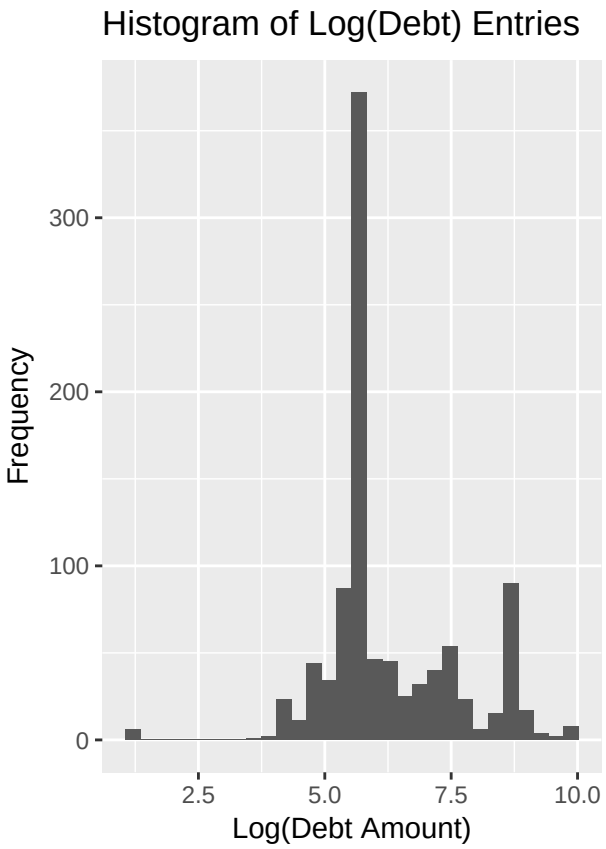
```
##  SEATTLE     :452   CA  : 10
##  WOODINVILLE :241   DC  :100
##  WASHINGTON  :100   TX  :  5
##  KIRKLAND    : 38   WA  :847
##  TACOMA      : 24   NA's: 25
##  (Other)     :108
##  NA's        : 24
```

Some interesting observations of debt when doing univariate metrics shows that there is a maximum of 19,000 dollars associated with consulting a service called "NEW PARTNERS CONSULTING INC" by Democrats and a minimum of 3.24 dollars associated with Republicans buying keys in the county of ISLAND. The median is also much less than the average meaning that the data is skewed to the right.

To visualize this, we generated a histogram of the debt filed for each entry in the dataset. The primary result (left) appears very skewed and perhaps even has an exponential shape. We then decided to do a log transformation (right) of the debt filed to get better granularity at the low end and a better look at any trends present.

```
plot1 <- ggplot(data=df, aes(df$amount)) +
  geom_histogram() +
  labs(title='Histogram of Debt Filed',
       x='Debt Amount',
       y='Frequency')
plot2 <- ggplot(data=df, aes(log(df$amount))) +
  geom_histogram() +
  labs(title='Histogram of Log(Debt) Entries',
       x='Log(Debt Amount)',
       y='Frequency')
grid.arrange(plot1, plot2, ncol=2)
```

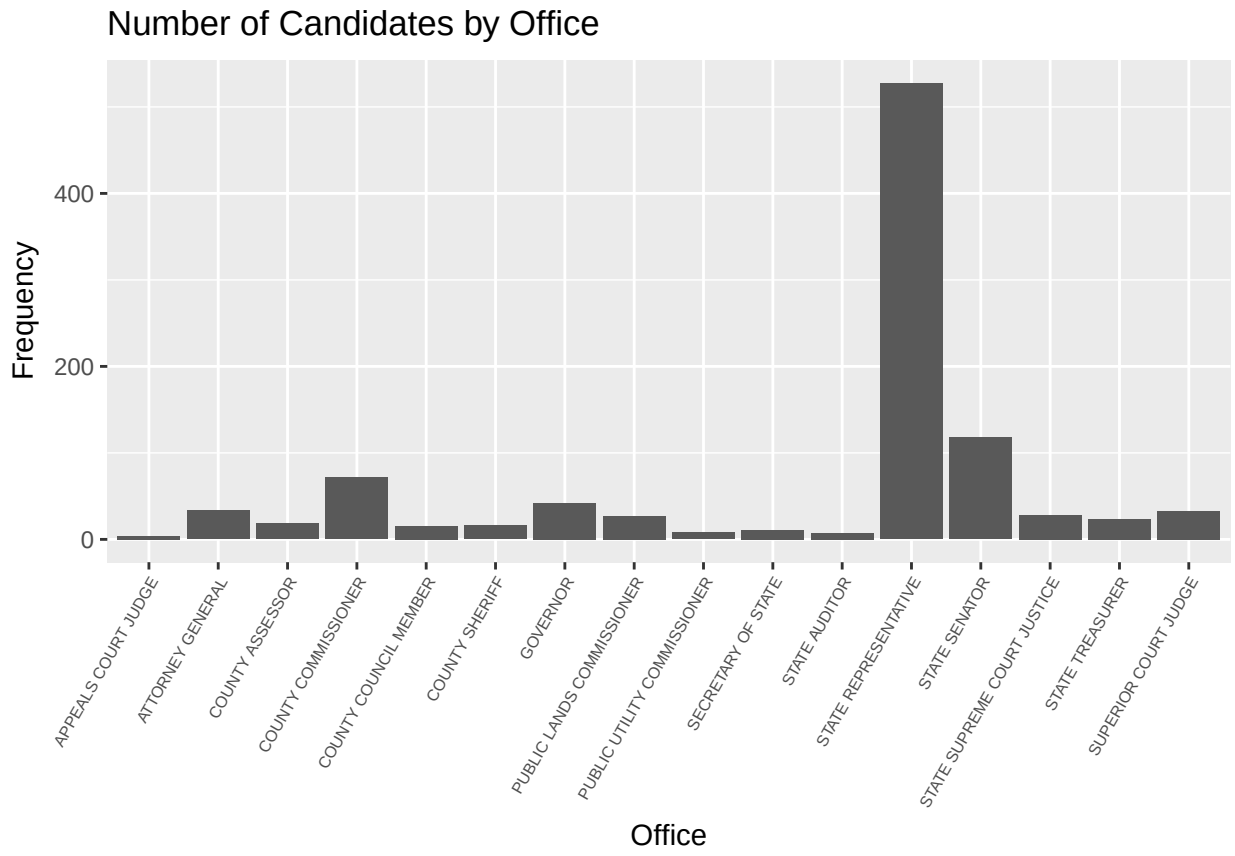




It appears that a majority of debt is incurred from lower cost spending across all candidates with an amount of approximately log(6) or $400 being the most frequent. It is hard to gain additional insight here as due to the clustering at the lower end of the distribution. We will revisit this as a function of other key variables in the multivariate analysis.

Next we wanted to look at the number of candidates in each position and office to get a sense of what offices the candidates were running for. To do this, we took a distinct count of each filerID per position and office.

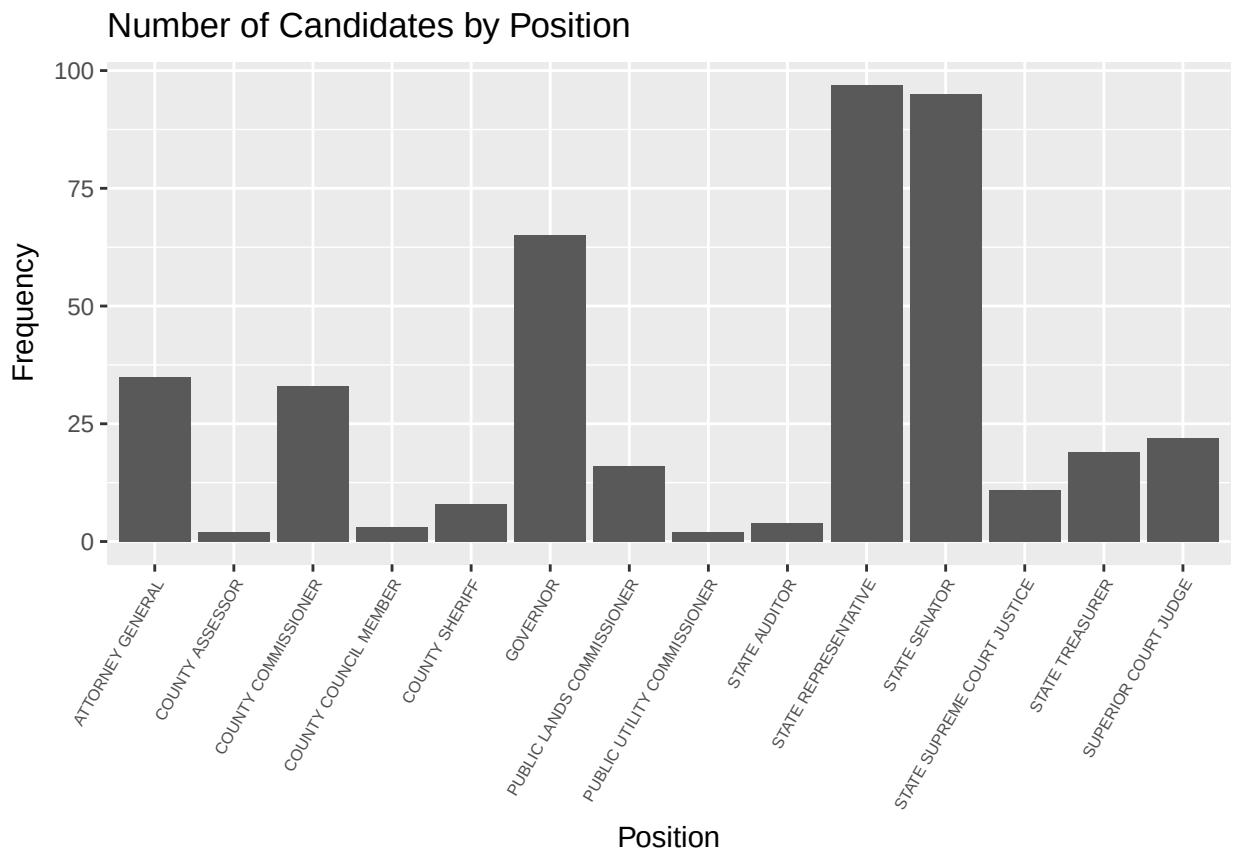
```
df_by_office = df %>%
  select(amount) %>%
  group_by(df$filerid, df$office) %>%
  summarize(sum(amount), n=n())
names(df_by_office) = c('filerid', 'office', 'amt_sum', 'amt_count')
ggplot(df_by_office, aes(office,amt_count)) +
  geom_bar(stat='identity') +
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(title="Number of Candidates by Office",
       x="Office",
       y="Frequency")
```



We observed there were more candidates for state representative than any other office by a large margin.

To get a general idea of which positions are the most frequently sought, we counted the number of candidates per position and found similar trends. State representative, state senator, and governor positions are the most popular.

```
df_filer_area <- unique(data.frame(df$filerid, df$position))
colnames(df_filer_area) <- c('filerid', 'position')

df_filer_area_data = sqldf(
    "select position, count(filerid) as count from df_filer_area group by 1"
  )

ggplot(data=df_filer_area_data, aes(x=position, y=count)) +
    geom_bar(stat="identity", position="dodge") +
    theme(axis.text.x = element_text(size=6, angle=60, hjust=1)) +
    xlab("Position") +
    ylab("Frequency") +
    ggtitle("Number of Candidates by Position")
```



The number of candidates per position and office will be used later during the multivariate analysis when we look at how debt interacts with this grouped data.

At this point we are ready to start inspecting the key variables in relationship to each other rather than in isolation.

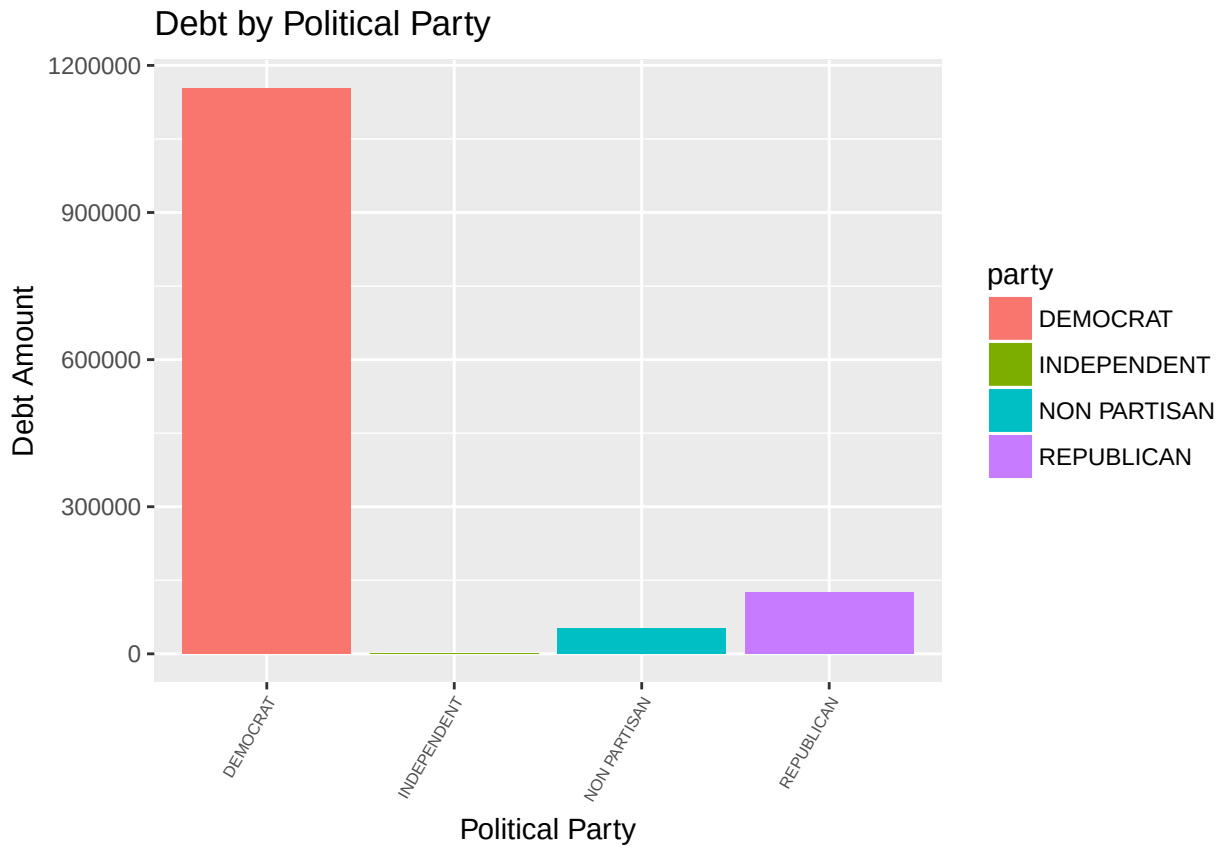
Multivariate Analysis of Key Relationships

To perform multivariate analysis, we decided to focus on mixing several of the variables described in the univariate analysis to see what additional insights could be gained. Throughout our analysis, we commonly break down the data by party.

Debt by Party

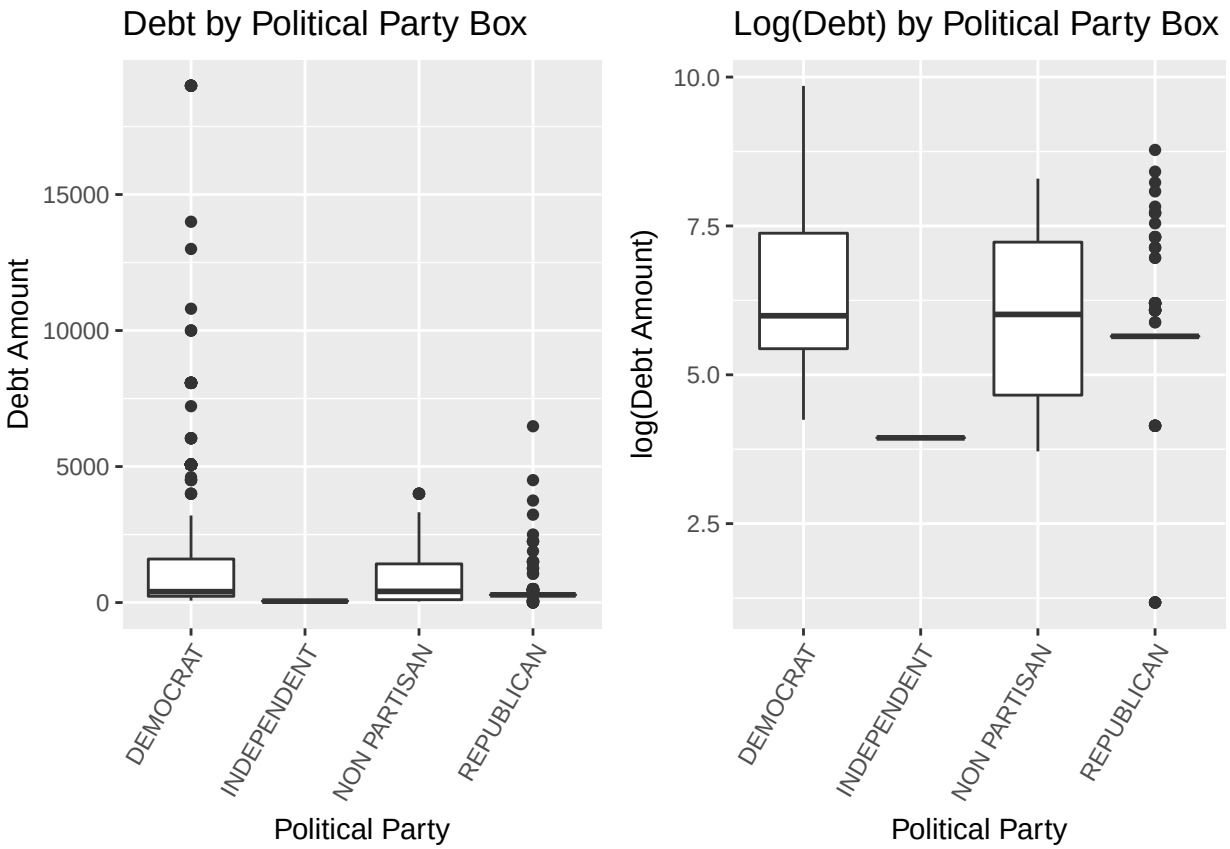
Since party will be a common theme in our analysis moving forward, it is important to firstly get an understanding of how debt and party relate to one another.

```
ggplot(data=df, aes(x=party, y=amount)) +
  geom_bar(stat="identity", aes(fill=party)) +
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(
    title='Debt by Political Party',
    x='Political Party',
    y='Debt Amount'
  )
```



Clearly, this shows that democrats are outspending all other parties in the election. To visualize the distribution of debt that make up the sums above, we created box plots of the debt by party. The one on the left was done without transformation but because it is hard to discern any trends, a log transformation was applied to give the box plot on the right.

```
df_party = df[!is.na(df$party), ]
plot1<-ggplot(df_party, aes(x=party, y=amount)) +
  geom_boxplot() +
  labs(
    title='Debt by Political Party Box',
    x='Political Party',
    y='Debt Amount'
  ) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))

plot2<-ggplot(df_party, aes(x=party, y=log(amount))) +
  geom_boxplot() +
  labs(
    title='Log(Debt) by Political Party Box',
    x='Political Party',
    y='log(Debt Amount)'
  ) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))

grid.arrange(plot1, plot2, ncol=2)
```

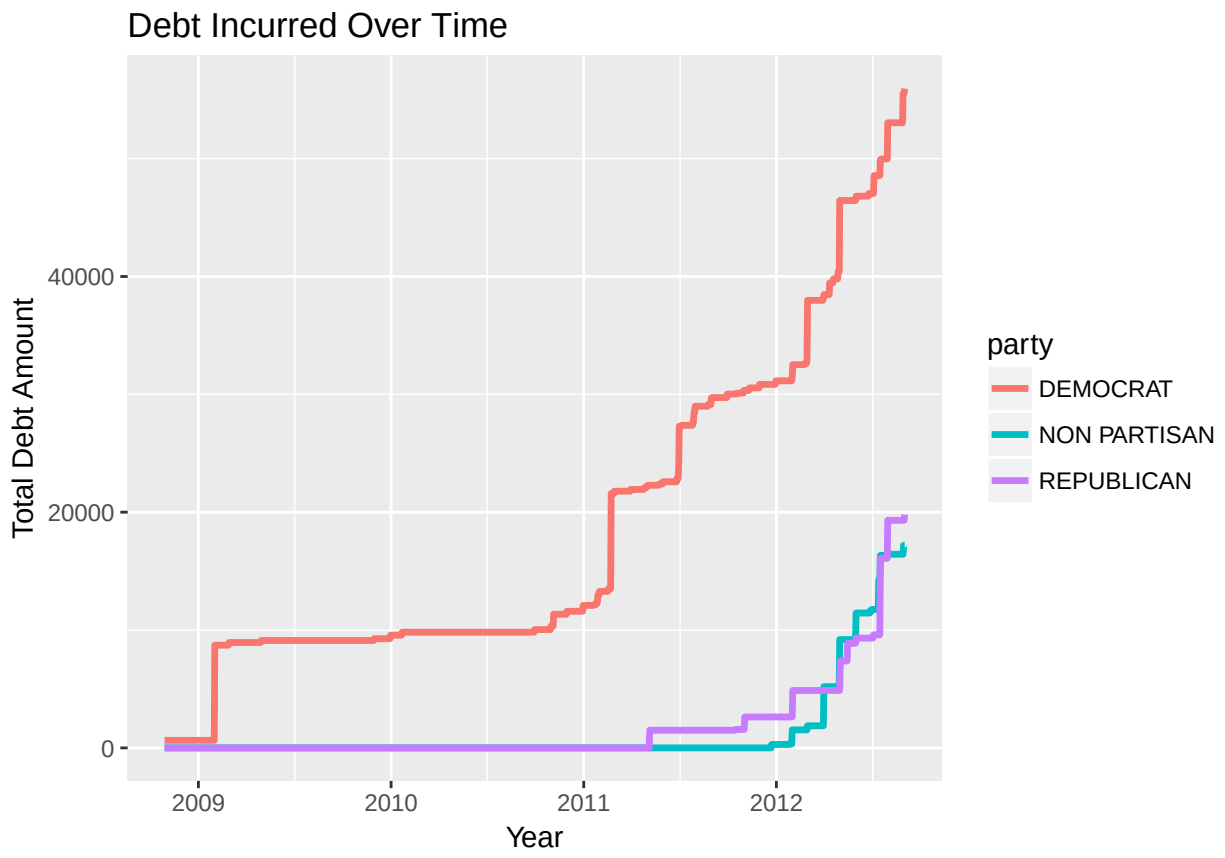


We see there is a wide distribution of debt amounts amongst the Republican and Democrat parties, but not as much in the Independent and Non-Partisan parties. This will be important to keep in mind as we further dissect how the debt interacts with other key variables that are grouped by party. Like before, there the data is very skewed towards higher debt but looking at the log transformed data it seems that Democrat and Non-partisan debt is very broad whereas Independent and Republican debt is very sharp and narrow. This causes a lot of outliers to appear for Republicans.

It was also important to get a sense of which party incurred the most debt year by year. The independent party was left out due to their low debt value.

```
df = df[!is.na(df$recordtype),]
df = df[colnames(df) != 'X']

df$debtdate <- as.Date(df$debtdate)
df_debtdate_all = as.data.frame(seq(min(df$debtdate), max(df$debtdate), by=1))
colnames(df_debtdate_all)[1] = 'debtdate'

df_democrat = sqldf("select a.debtdate, ifnull(party,'DEMOCRAT') as party,
                    ifnull(amount,0) as amount from df_debtdate_all a
                    LEFT JOIN (select debtdate,party, amount from df
                    where party = 'DEMOCRAT' group by 1,2)
                    b on a.debtdate = b.debtdate
                    ")

df_democrat[, 3] <- cumsum(df_democrat[, 3])

df_republican = sqldf("select a.debtdate, ifnull(party,'REPUBLICAN') as party,
```

```
                        ifnull(amount,0) as amount from df_debtdate_all a
                        LEFT JOIN (select debtdate,party, amount from df
                        where party = 'REPUBLICAN' group by 1,2)
                        b on a.debtdate = b.debtdate
                       ")

df_republican[, 3] <- cumsum(df_republican[, 3])

df_nonp = sqldf("select a.debtdate, ifnull(party,'NON PARTISAN') as party,
                 ifnull(amount,0) as amount from df_debtdate_all a
                 LEFT JOIN (select debtdate,party, amount from df
                 where party = 'NON PARTISAN' group by 1,2)
                 b on a.debtdate = b.debtdate
                ")

df_nonp[, 3] <- cumsum(df_nonp[, 3])

df_debt_both_parties = bind_rows(df_democrat, df_republican, df_nonp)

ggplot(data=df_debt_both_parties, aes(x=debtdate, y=amount, group=party)) +
geom_line(size=1.2, (aes(color=party))) +
xlab("Year") +
ylab("Total Debt Amount") +
ggtitle("Debt Incurred Over Time") +
scale_colour_manual(values = c("DEMOCRAT" = "#f8766d",
                               "REPUBLICAN" = "#c77cff",
                               "NON PARTISAN" = "#00bfc4",
                               "INDEPENDENT" = "#7cae00"))
```



Plotting the debt by year and party shows us that the Democrats accumulated significantly more debt than any other party. It can readily be seen that the Democrats acted early and strongly, investing a lot for the 2012 election as early as 2009. They then started investing heavily in 2011 spending more than double their investment total prior to 2011. Conversely, both Republicans and the Non-partisan party don't put significant investment into anything until the latter half of 2011 / beginning of 2012. Both groups also invested less than half the amount compared to Democrats by the end of the election. Looking at the trend after parties seem to start campaigning seriously, it almost looks linear. This indicates that there could be potential for prediction in future downstream analysis.

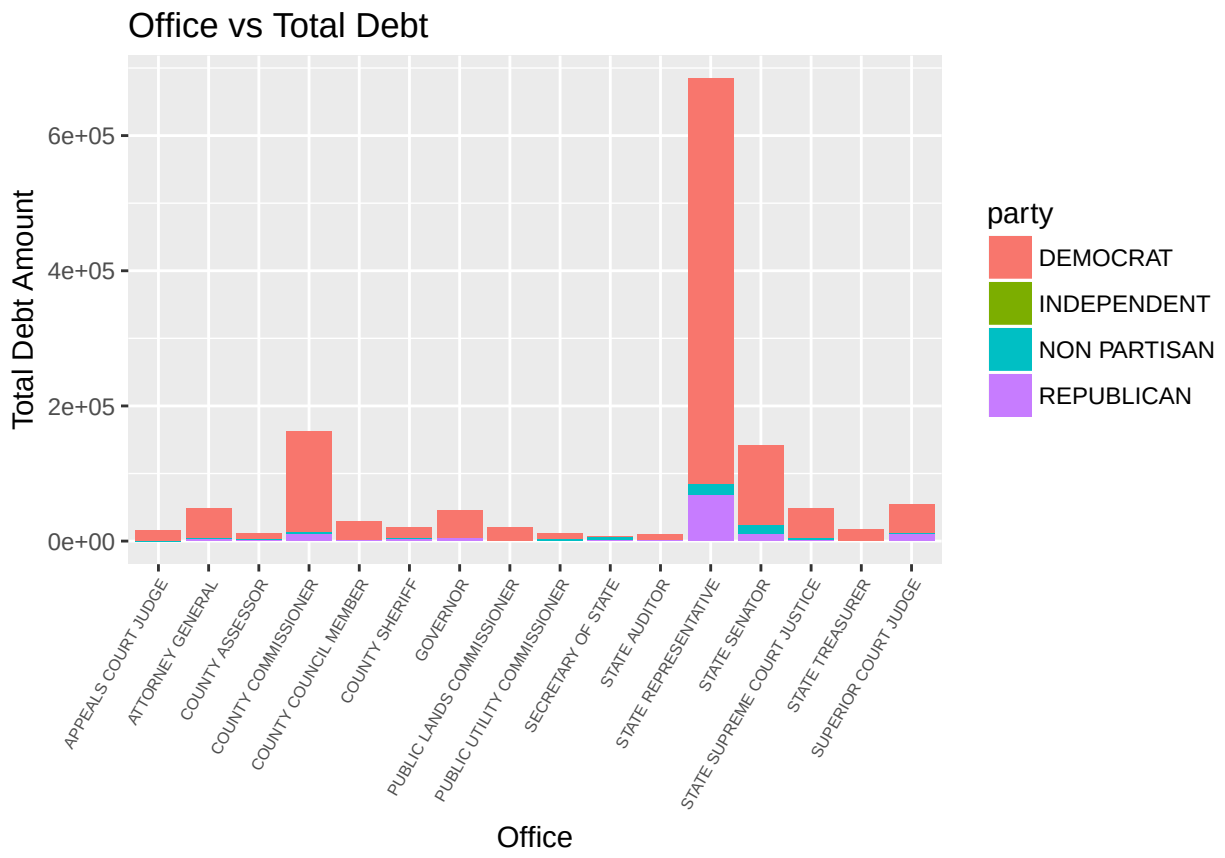
Relationships of Debt to Office

The second relationship we want to explore is how the debt incurred by the candidate relates to office and party.

```
print(length(unique(df$filerid)))
```

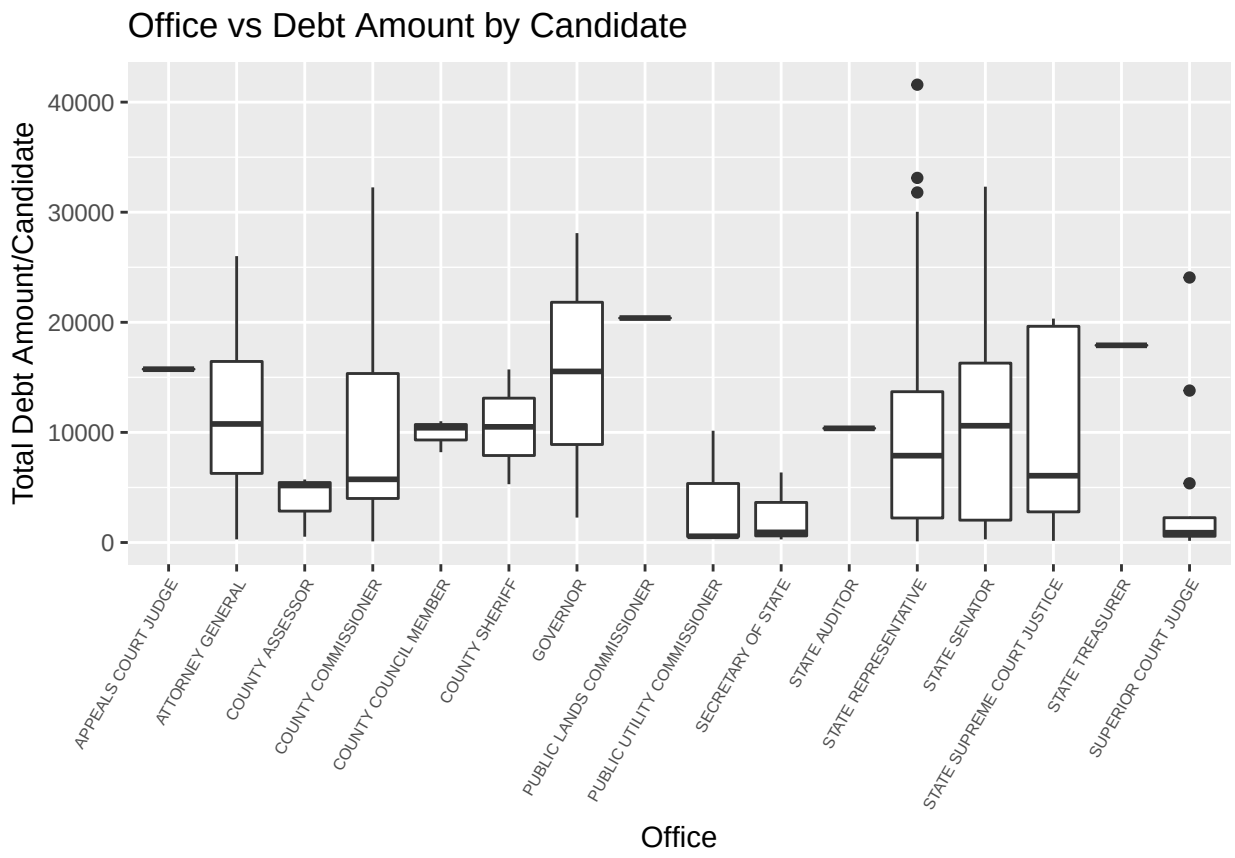
```
## [1] 140
```

```
ggplot(df, aes(office,amount)) +
  geom_bar(stat='identity', aes(fill=party)) +
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(title="Office vs Total Debt", x='Office', y='Total Debt Amount')
```



One immediate observation is that Democrat spending dominates all offices in the Washington area. To increase granularity, we look at the data on a per candidate level by grouping by filerID and and then seeing the spread of the total debt amount by candidate versus which office the candidates were running for.

```
ggplot(df_by_office, aes(office,amt_sum)) +
  geom_boxplot() +
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(
    title="Office vs Debt Amount by Candidate",
    x='Office',
    y='Total Debt Amount/Candidate'
  )
```



In general, this plot shows us the median and interquartile range of the debt filed by candidate for each office. From this we can see that the candidates for Governor's office has the largest median debt, but the State Supreme Court Justice has the largest variation given by its interquartile range. We also examined the offices where it seemed like the ranges were very tight (State Treasurer, State Auditor, Public Lands Commissioner, Appeals Court Judge) and noted that for those offices, there was effectively only one entry so any statistical conclusions on these offices would be difficult.

```
df_by_office[df_by_office$office=='STATE TREASURER' |
             df_by_office$office=='PUBLIC LANDS COMMISSIONER' |
             df_by_office$office=='STATE AUDITOR' |
             df_by_office$office=='APPEALS COURT JUDGE',]
```

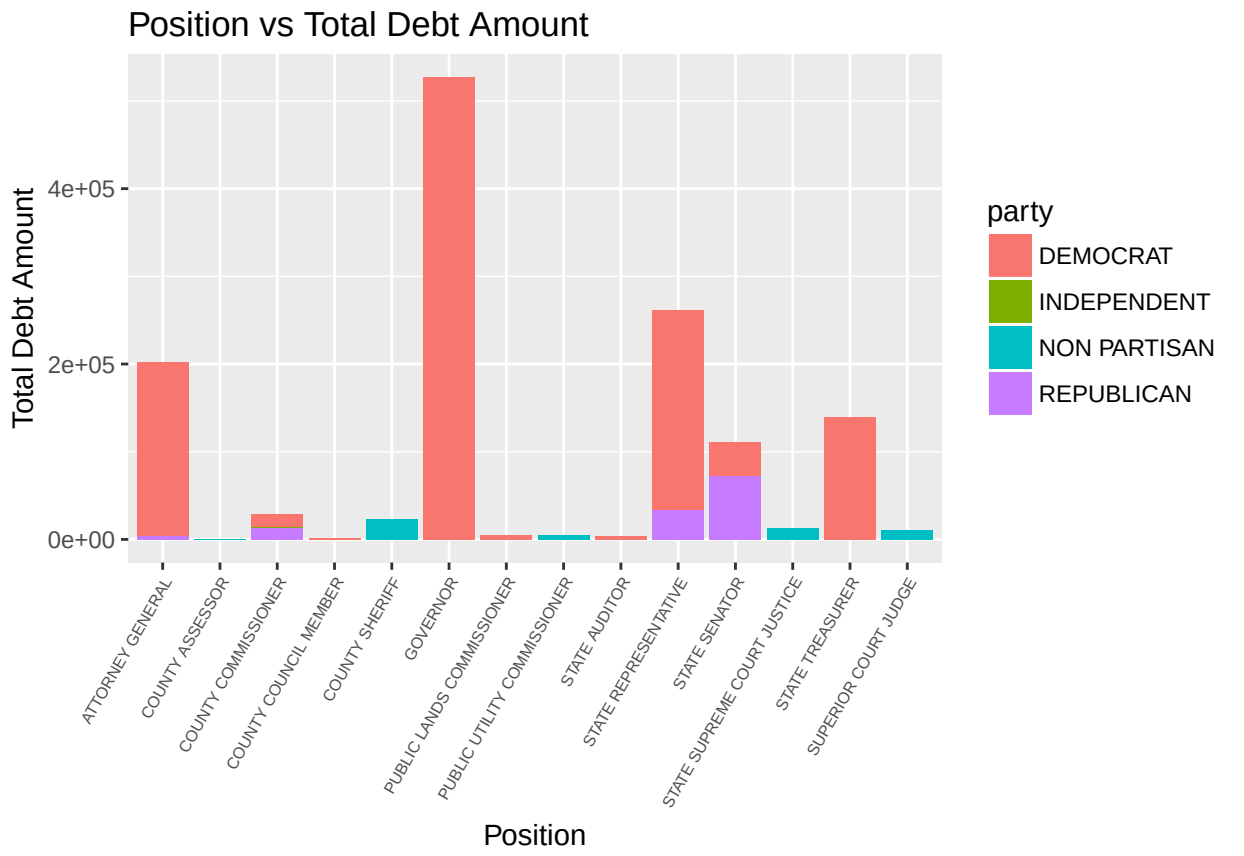
```
## # A tibble: 4 x 4
##   filerid    office                     amt_sum amt_count
##   <fct>      <fct>                        <dbl>     <int>
## 1 GOLDP  840 PUBLIC LANDS COMMISSIONER  20387.        27
## 2 KELLT2 406 STATE AUDITOR               10369.         7
## 3 MCINJ  115 STATE TREASURER             17910.        24
## 4 WILLB2 501 APPEALS COURT JUDGE         15742.         4
```

Relationship of Debt to Position

We then observed the total debt as a function of position, grouped by party.

```
ggplot(df, aes(position, amount)) +
  geom_bar(stat='identity', aes(fill=party)) +
```

```
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(title="Position vs Total Debt Amount",
       x='Position',
       y='Total Debt Amount')
```



From this we can see that Democrats spend the most on the Governor, State Representative, and Attorney General position. Spending for the Governor position is by far the the greatest and this is consistent with the result above when looking at debt per candidate by office.

Also, although heavily sought after though, this result on Governor is a bit deceiving in the sense that almost all debt appears to be collected by Democrats whereas the Office vs Total Debt graph does show a bit of Republican presence in the data.

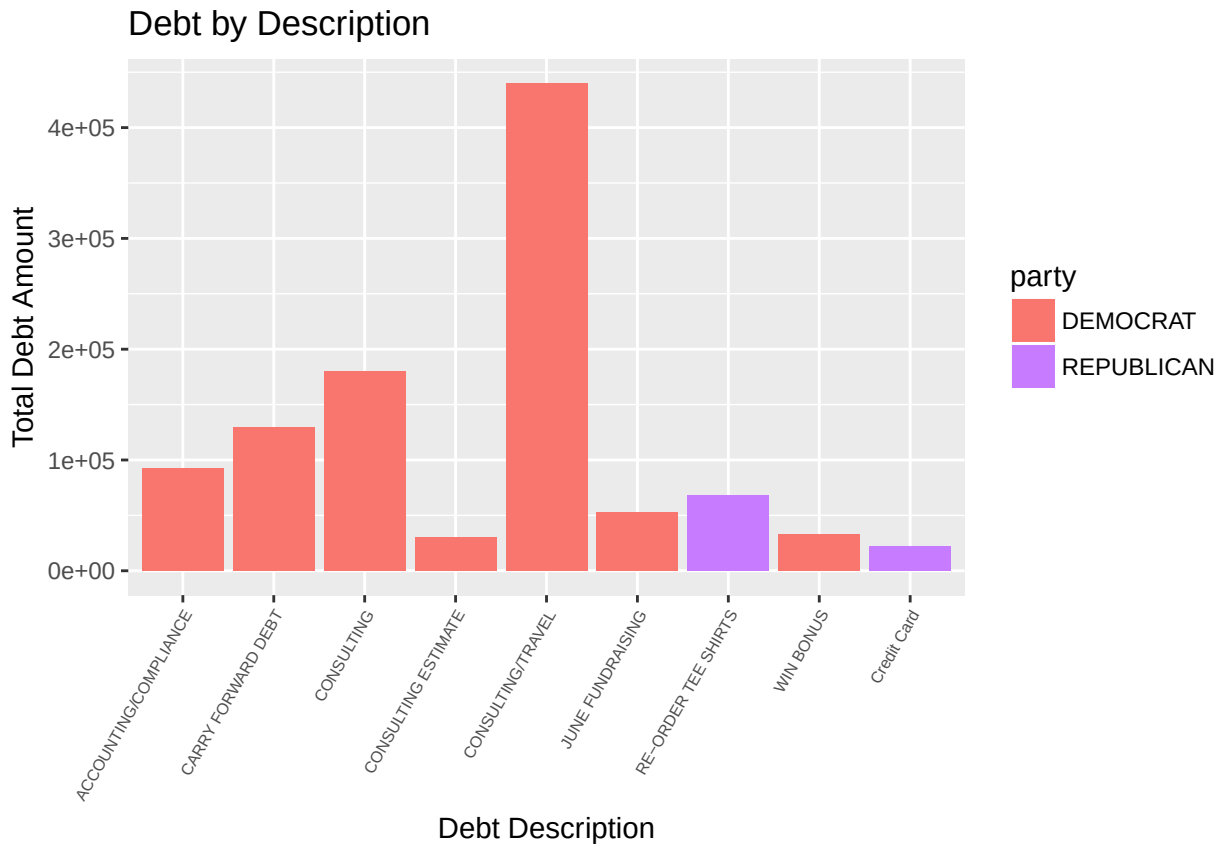
Lastly it's important to remark that State Senator is an interesting result. It is the only position where Republicans have more spending than Democrats and as remarked before, Republicans spend very little. This is inverted when looking at the data from a debt vs office perspective. We find that in the data this is because the office of State Senator contains many individual positions. This suggests that Republicans have a very focused strategy of getting the State Senator position itself..

Relationship of Debt to Vendors and Services Spent On

Now we want to see what these candidates are spending all this money on. First we look at the total about of debt versus the description of the transaction.

```
df_party_description_sum <- data.frame(df$description, df$party, df$amount)
colnames(df_party_description_sum) <- c('description', 'party', 'amount')
```

```
df_party_description_sum_data = sqldf(
  "select description, party, sum(amount) as debtamount
  from df_party_description_sum group by 1,2 having debtamount > 20000"
)

ggplot(data=df_party_description_sum_data, aes(x=description, y=debtamount)) +
  geom_bar(stat="identity", aes(fill=party)) +
  theme(axis.text.x = element_text(size=6, angle=60, hjust=1)) +
  xlab("Debt Description") +
  ylab("Total Debt Amount") +
  ggtitle("Debt by Description") +
  scale_fill_manual(values = c("DEMOCRAT" = "#f8766d",
                               "REPUBLICAN" = "#c77cff",
                               "NON PARTISAN" = "#00bfc4",
                               "INDEPENDENT" = "#7cae00"))
```



Consulting/Travel is the major source of debt. One interesting observation is the 'Carry Forward Debt' seems quite large from prior elections. We can also see that the 'Credit Card' identifier we created previously was a common transaction and was the description for a significant amount of debt (which is not unexpected for an election taking place in 2012)

Another interesting observation comes from looking at the individual vendors themselves.

```
df_by_vendorname=aggregate(df$amount,by=list(df$vendorname),sum)
names(df_by_vendorname)=c('vendorname','amt_sum')
head(df_by_vendorname[order(df_by_vendorname$amt_sum,decreasing = T),], 5)
```

```
##                       vendorname   amt_sum
## 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
```

As expected, consulting companies are some of the highest costing vendors. However, there is also a non-company listing, James McIntire.

```
head(df[df$vendorname=='MCINTIRE JAMES' & df$description=='CARRY FORWARD DEBT',
        c('filername', 'party', 'description', 'vendorname')], 20)
```

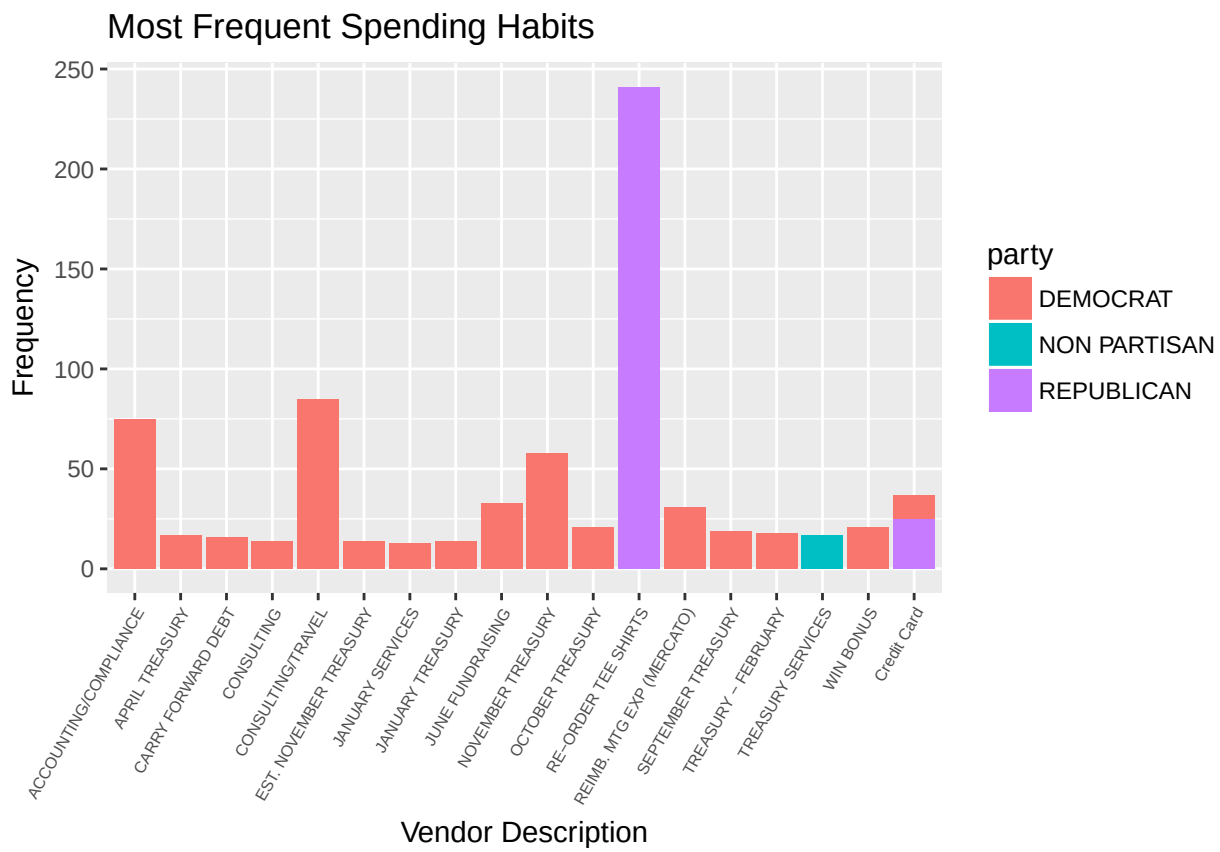
```
##                filername    party        description     vendorname
## 350     MCINTIRE JAMES L DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 354         INSLEE JAY R DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 355       FROCKT DAVID S DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 374         BROWN LISA J DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 375  LADENBURG JOHN W SR DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 377        CODY EILEEN L DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 381 PRENTICE MARGARITA L DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 383     MORRIS JEFFREY R DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 384  SPRINGER LAWRENCE S DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 386 GREGOIRE CHRISTINE O DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 390   SULLIVAN PATRICK J DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 391  VAN DE WEGE KEVIN W DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 411    CLIBBORN JUDITH R DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 412   ROLFES CHRISTINE N DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 440       DUNSHEE HANS M DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
## 441  SPRINGER LAWRENCE S DEMOCRAT CARRY FORWARD DEBT MCINTIRE JAMES
```

It appears that James McIntire focuses exclusively on funding the Democrats debt from prior elections, but it is very interesting that it also appears that he is running as a candidate for State Treasurer. This implies that Democrats might be spending state money to fund their debt. The fact that James McIntire might have the ability to be in debt to himself, or that his fellow party members might have the same opportunity brings up many ideas of conflict of interest and this is certainly something that deserves further investigation.

Next we want to look at the frequency of the spending habits for each party. What does each party spend money on? To investigate this, we counted the different descriptions of spending in the data ignoring those transactions with a count of less than 10 instances.

```
df_party_description <- data.frame(df$description, df$party)
colnames(df_party_description) <- c('description', 'party')
df_party_description_data = sqldf(
    "select description, party, count(description) as count from
    df_party_description group by 1,2 having count > 10"
  )

ggplot(data=df_party_description_data, aes(x=description, y=count)) +
  geom_bar(stat="identity", aes(fill=party)) +
  theme(axis.text.x = element_text(size=6, angle = 60, hjust = 1)) +
  xlab("Vendor Description") +
  ylab("Frequency") +
  ggtitle("Most Frequent Spending Habits") +
  scale_fill_manual(values = c("DEMOCRAT" = "#f8766d",
                               "REPUBLICAN" = "#c77cff",
                               "NON PARTISAN" = "#00bfc4",
```

```
                                "INDEPENDENT" = "#7cae00"))
```

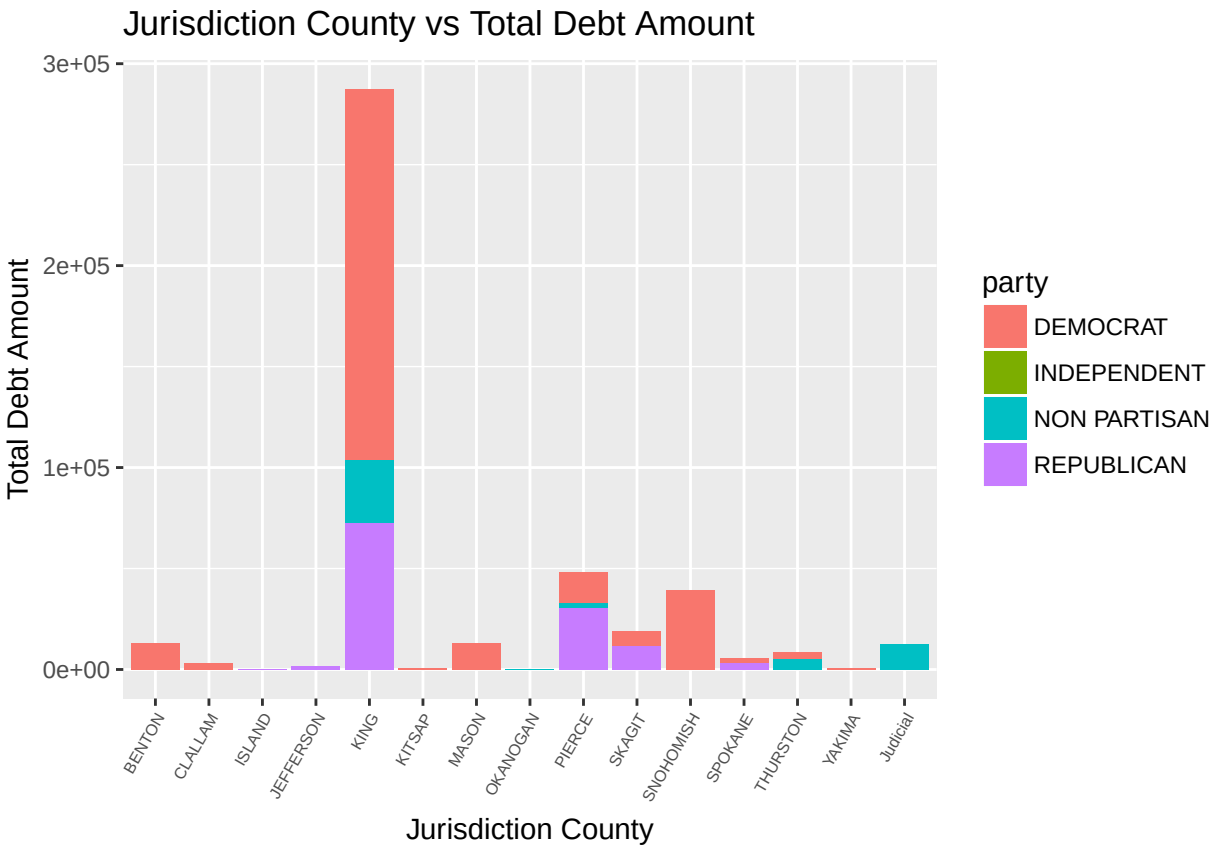


We find that Republicans place a lot of their funds in ordering t-shirts whereas Democratic spending is a lot more diverse. The largest service that Democrats spend on seems to be consulting, travel, accounting, and compliance. This observation of all parties speaks greatly to their individual strategies. It seems the Democrats in Washington state develop deep strategies on how to run their campaign while Republicans seem to focus more on local and word-of-mouth strategies. Another interesting observation is that with credit cards, it appears that both Democrats and Republicans used credit cards but only the Republicans incurred any debt.

Relationship of Debt to Jurisdisction

Here we look at the relationship of debt to which jurisdiction the candidate is running in. We remove the 'statewide' entries to get a better sense of the distribution in the smaller counties

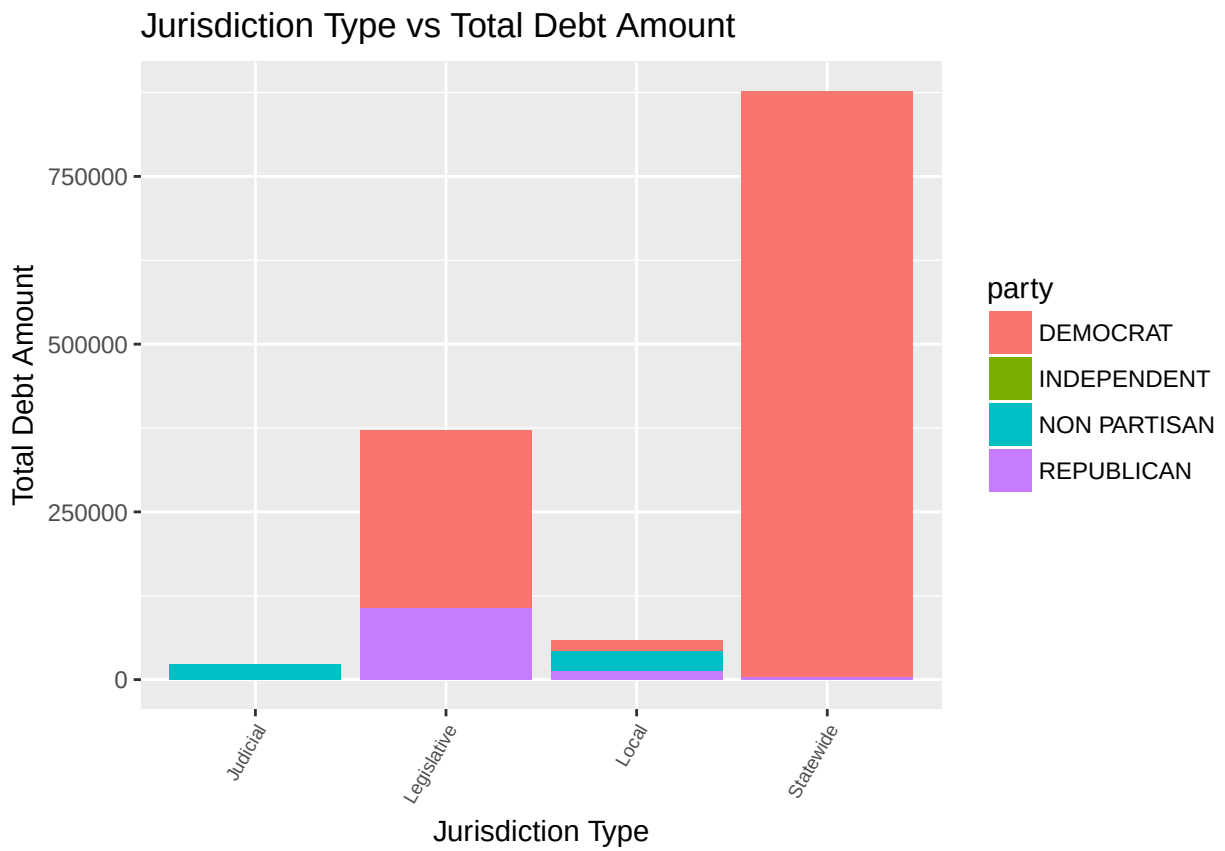
```
df_countyOnly = df[df$jurisdictioncounty != 'Statewide', ]
ggplot(df_countyOnly, aes(jurisdictioncounty, amount)) +
  geom_bar(stat='identity', aes(fill=party)) +
  theme(axis.text.x=element_text(size=6, angle=60, hjust=1)) +
  labs(title="Jurisdiction County vs Total Debt Amount",
       x='Jurisdiction County',
       y='Total Debt Amount')
```



Jurisdiction county appears to be a good indicator as to whether aggregate debt is likely to be high for a given party. King county, (legislative district 1), and Pierce county (legislative district 28) appear to be only counties with significant republican debt. It is interesting to note that Pierce and Skagit counties are the only counties where Republican debt was greater than Democrats' debt. This is perhaps an indication that Republicans really want to win these areas.

Next we want to look at the breakdown of debt versus jurisdiction type

```
ggplot(df, aes(jurisdictiontype, amount)) +
  geom_bar(stat='identity', aes(fill=party)) +
  theme(axis.text.x=element_text(size=7, angle=60, hjust=1)) +
  labs(title="Jurisdiction Type vs Total Debt Amount",
       x='Jurisdiction Type',
       y='Total Debt Amount')
```



We see that the legislative and statewide jurisdictions accumulate the most debt which is not surprising. We also see that the judicial jurisdictions are non-partisan which is also expected since most states use a non-partisan method to select their judges.

Analysis of Secondary Effects

Confounding variables

The 'filerid' variable is the unique identifier given to each candidate upon filing. We noticed that there were 140 such unique entries.

```
length(unique(df$filerid))
```

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

Based on this all "simple" factors should have 140 obs

```
dim(df %>% select(amount) %>%
      group_by(df$filerid, df$position) %>%
      summarize(sum(amount), n = n())
  )[1]
```

```
## [1] 412
```

```
dim(df %>% select(amount) %>%
      group_by(df$filerid, df$party) %>%
      summarize(sum(amount), n = n())
```

```
  )[1]
```

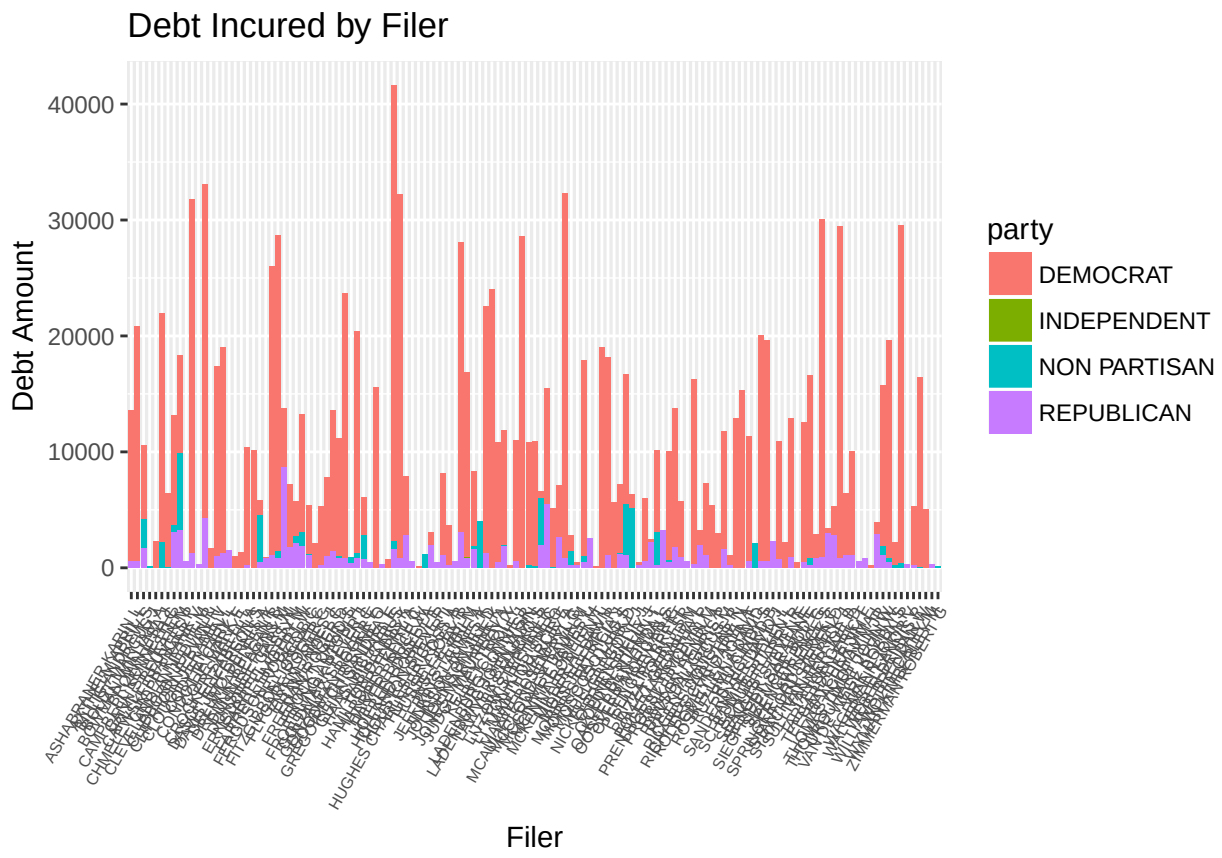
```
## [1] 257
```

```
dim(df %>% select(amount) %>%
      group_by(df$filerid, df$jurisdictiontype) %>%
      summarize(sum(amount), n = n())
  )[1]
```

```
## [1] 288
```

This observation shows that candidates are switching parties and jurisdictions and running for multiple positions. This behavior almost certainly is impacting many of the relationships that we observed prior, but an in-depth analysis of the impact this switching behavior is having is beyond the scope of this EDA. We can at least see if this behavior is rampant throughout all candidates, or isolated to just a few.

```
ggplot(data=df, aes(x=df$filername, y=amount)) +
  geom_bar(stat="identity", aes(fill = party)) +
  theme(axis.text.x = element_text(size=6, angle = 60, hjust = 1)) +
  xlab("Filer") +
  ylab("Debt Amount") +
  ggtitle("Debt Incured by Filer")
```



This graph shows the debt filed by filername and is grouped by party. Though it has a cluttered x-axis, the idea is to show that some bars have more than one color. Therefore, it is easy to see that many of the candidates are switching parties mid election.

Conclusions

Initial Observations

Since the key variables are almost all categorical factors, there is almost no transformation of the independent variables. The max debt seems to be 19,000 dollars, spent by Democrats on consulting services whereas the min seems to be 3.24 dollars which were spent on keys by Republicans. It is not known why the spending is capped at 19,000 dollars but likely it could be there is some legislative filing limit.

Key Variables

Office

Most of the money was spent by candidates running for State Representative. It was almost 3x larger than the next closest office - county commissioner - in aggregate.

Looking at the total debt per candidate and by office, we can see that the candidate for office of governor got into much more debt on average. It is interesting to note that other large debt accumulations by office - public lands commissioner, state treasurer and appeals court judge - were only to have 1 candidate in debt - and they got in quite a bit of debt as well.

Position

Position (what "bigger" role does the office sit under), in aggregate, is dominated by positions in the Governor's Office with the largest debt accumulation. It is interesting to note that Positions under State Senator is one of the few where Republicans got in more debt than democrats. Again, this speaks to their strategy of pursuing that position in particular.

Jurisdiction

The jurisdiction appears to be a big factor for spending depending on the party (Republican or Democrat). We observed that Statewide elections accumulate most debt as expected. However, in districts, 1 - SENATE and 28 - HOUSE, Republicans accumulated much more debt than democrats with almost no debt relatively elsewhere. Jurisdiction appears to be an interesting variable for Republican debt.

Jurisdiction County

Jurisdiction county appears to be a good indicator as to whether aggregate debt is likely to be high for a given party. King county, (legislative district 1), and Pierce county (legislative district 28) appear to be only counties with significant republican debt. It is interesting to note that Pierce and Skagit counties are the only counties where Republican debt was greater than Democrats' debt.

Debt Date

If modeling by how much debt is accumulated through out the years to election, it can be a good predictor of debt - especially given the party of candidate. This can be seen by the linearity of the trend once campaigning ramps up.

Party

Dominated almost by Democrats. This is possibly due to the fact that the data is from a state that is a traditionally blue, but it could be a potential predictor of how much debt a candidate will take on.

Vendor

From the analysis, there were two interesting observations. One is that although Republicans don't incur debt as much as Democrats, they are very inclined to buy t-shirts. This gives us a better sense of their strategy. Another interesting finding was that James McIntire appears to be running for State Treasurer and is the vendor for "Carry Over Debt" for all of the Democrats who have this type of debt. Without understanding the laws surrounding this, at a glance this appears to be illegitimate and is certainly something worth further exploration.

Confounding Variables/Secondary effects -

It appears that many candidates are switching parties, offices, positions, and jurisdictions throughout the election cycle. This effect undoubtedly has an impact on how we look at the data and draw conclusions since this phenomenon is happening at a very granular level. Although, it is difficult to do any picking for a distinct value since this could be considered biased data manipulation, it also presents an opportunity to further use this dataset given enough clarification is sought after.

Overall Summary

Despite some confounding observations, we can still gather some fairly interesting observations and conclusions. Washington is clearly a Democrat state. They are aggressive terms of incurring the most debt for goods and services for almost every position available in government. Much of their spending was on expensive consulting services. They definitely wanted a large presence in the politics of the state. Conversely, Republicans don't spend as much money but they have a very defined strategy. They spend the most on T-shirts which they probably see as an affordable option for winning. They also have the most debt in 2-3 jurisdictions only and are focused on getting their representatives to win the State Senator position, which is definitely a highly influential one.