Yanyan Liu

I only looked at 2016 contract data.

- **Number of initial records.**
  There are totally 4802211 records.

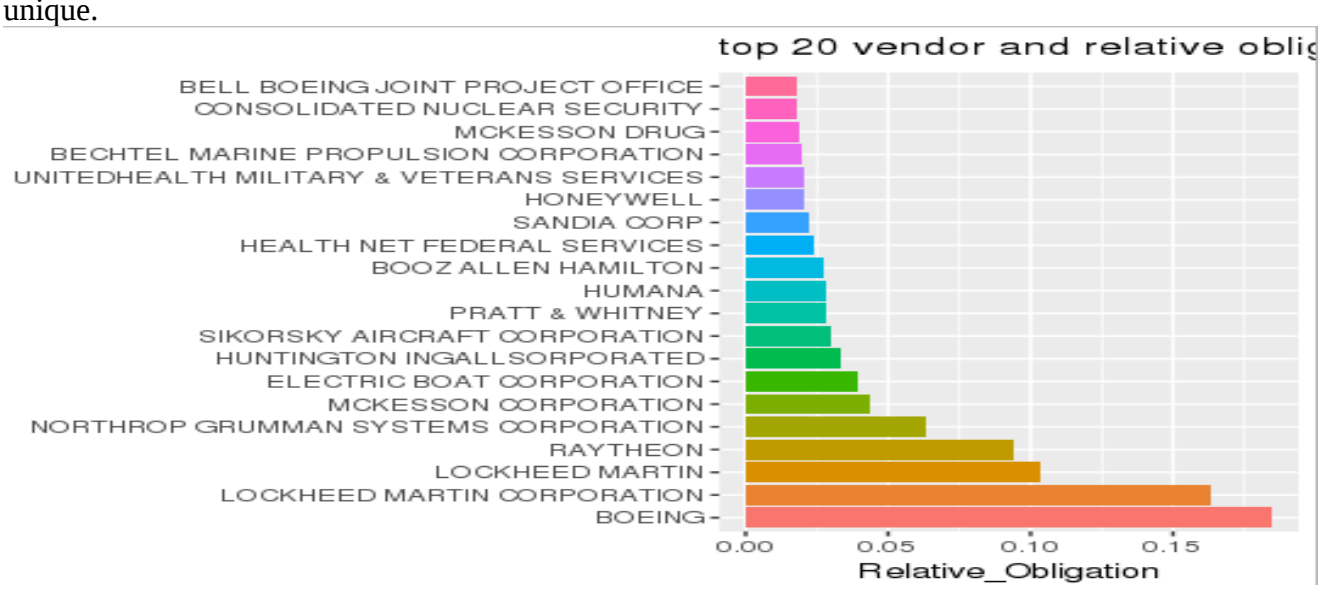
- **Number of records after reducing based on exact supplier/vendor matches.**
  After exact supplier/vendor matches, it shows there are 152597 "unique" vendors.
  Based on the requirement, I selected the rows with obligated money bigger than 25000, there are 636621 records. Then using exact supplier/vendor match, there are 66015 "unique" vendors.

- **Number of records after reducing further based on "fuzzy" matching criteria. This should group together records where the same supplier had slightly different names (such as "W.W. Grainger" and "WW Grainger" or "IBM" and "International Business Machines"). Some of the fuzzy matching logic might also mean matching across columns such as matching vendorname with vendoralternatename. Fields like phonenumber, streetaddress, city, state, and dunsnumber can also provide useful signals.**

I looked at vendorname and vendoralternatename column and find that vendoralteratename column seems to be more concise, so I compared the string length of names from these two columns, after some string manipulation, (change into capital letters, get rid of special characters like dot).
I suppose there must be some string distance algorithm can be used to do more detailed fuzzy matching, but I did not implement this.
There are 64798 "unique" vendors after this.
There are 64921 "unique" dunsnumber after this.

- **Some measure(s) of accuracy with explanations.**

1) I graphed the vendors with top 20 total obligated-money amounts; all top 20 vendor's name is unique.

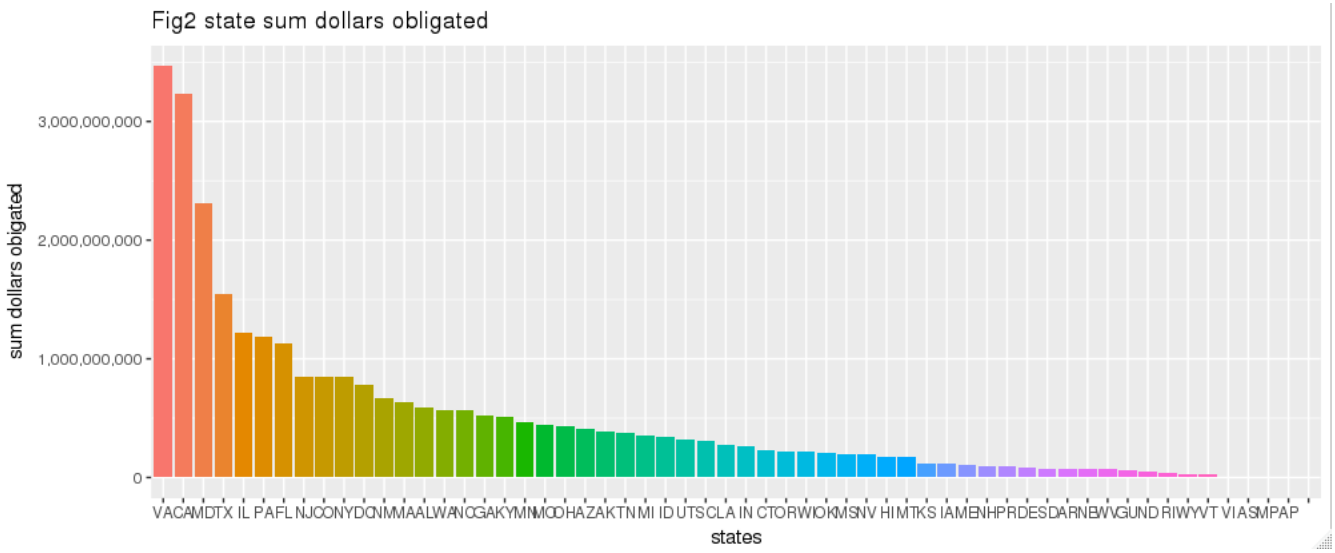


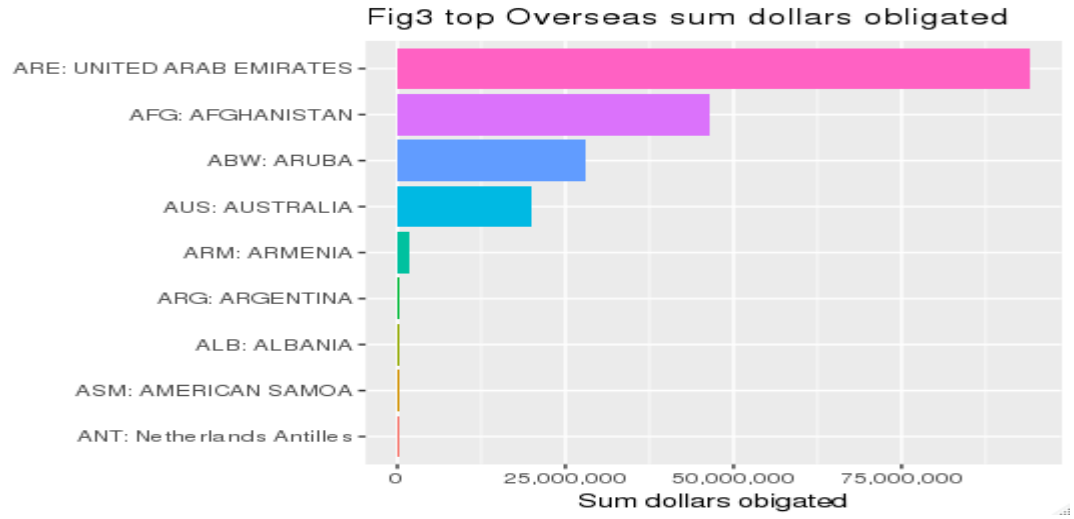
2) the unique count of dunsnumber is very close to the unique vendor count as is stated above.

Once you have grouped together suppliers what questions does this enable you to answer that you couldn't answer as well/accurately before performing this curation?
Code see the R file.
As the above, if we did not group with the vendor, it is hard to see the summary data for the top20

vendors.

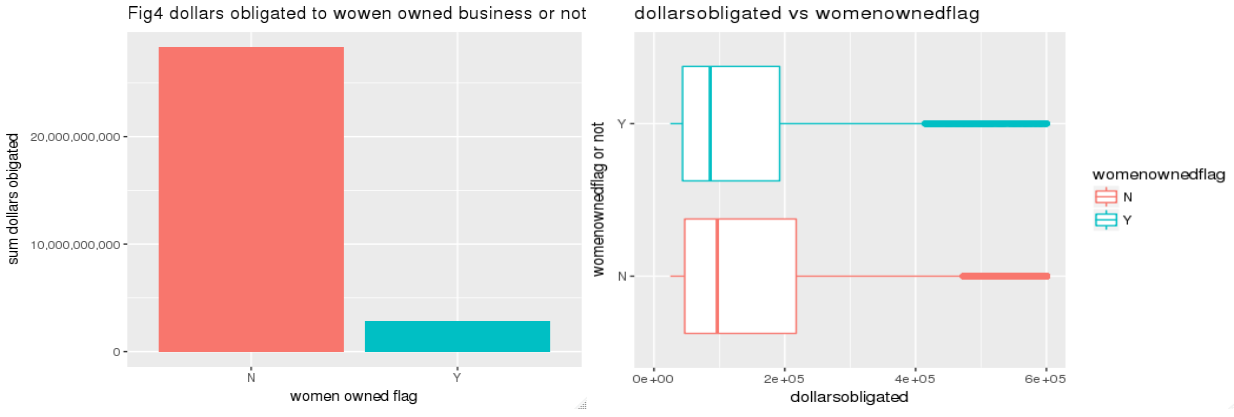
Just for curiosity, I also plotted the obligated money by state. Of course, to see this, it does not need the cleaned data to see it.

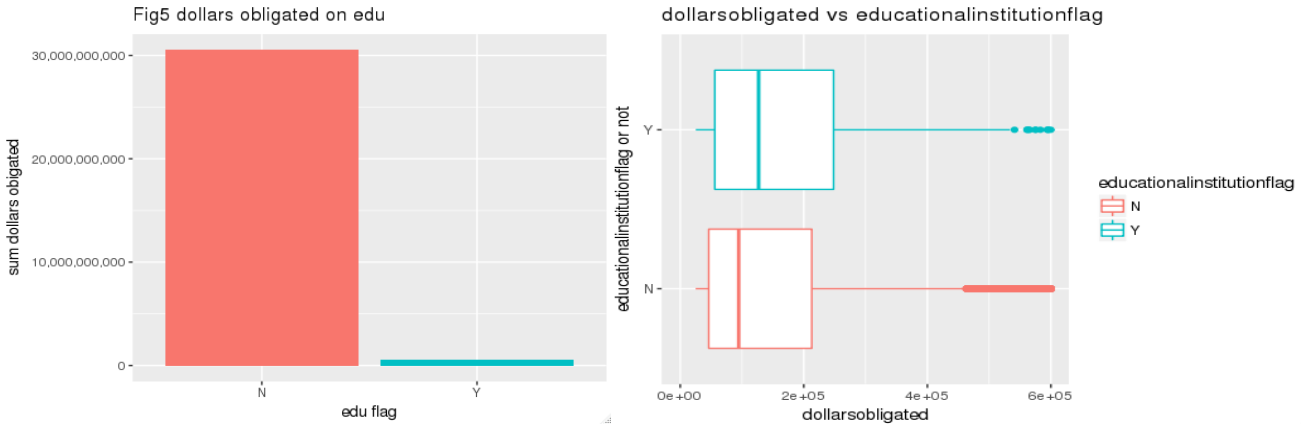




The above is the spending overseas. We see there are 4 top oversea countries that got money from USA government.

I also plotted the obligated money using women owned flag and education flag, which does not need the data clean also. The government spending on vendors owned by women has relative small percentage.
The government spending amount on vendors that have education flag is also relatively small.

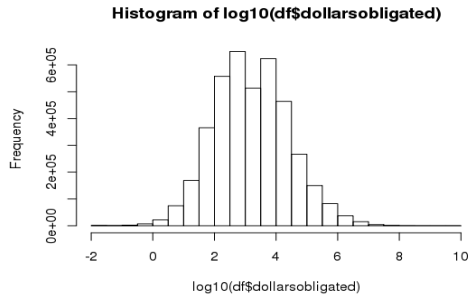




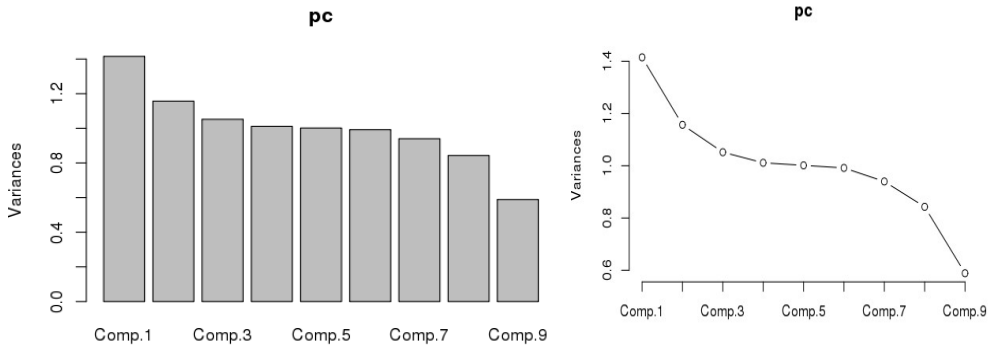
The script you used (python, Hive, or otherwise) in order to generate clusters of suppliers. Any other interesting results/conclusions.

Interesting results:

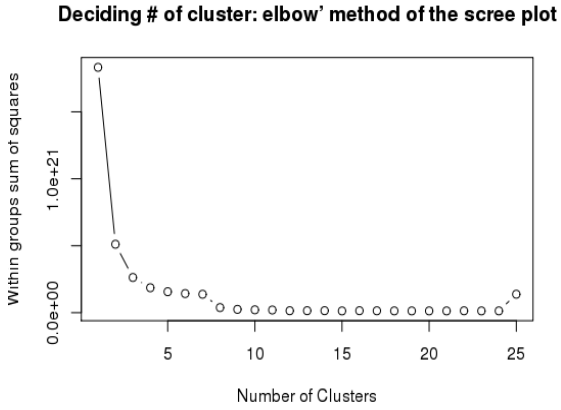
1) The money obligated is skewed toward right, log transformation is needed to plot the histogram of the money obligated. Not related here, just interesting to show.

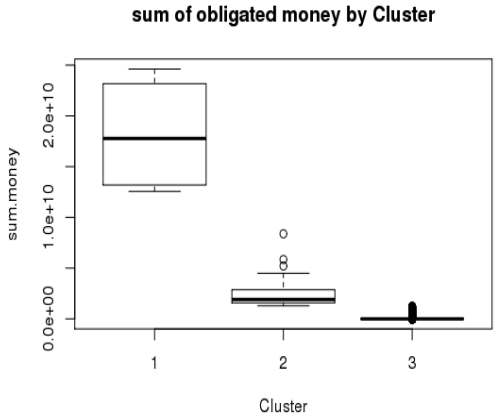


2) Which variable to choose relies on domain knowledge, I selected the various flag (countrycode, education, government, education, non-profit, emergency, hospital) and the total obligation, totaling 9 rows to do the clustering. By PCA analysis, we can see all 9 components contribute to the variance, even the smallest one is not negligible.



I did scree plot to decide the K to be used in the clustering, and found the K equals 3 seems to be a good number, as after that sum of square errors did not change much.





When clustering these vendors into 3 clusters, we can see clear sum of obligated money difference between these 3 clusters. We can see the obligated money played a significant role in forming these three clusters. If I start with different variables, I might come with different cluster.