Analysis of City of Chicago Payment Dataset

The data file and dataset description can be found at [City of Chicago](https://data.cityofchicago.org/Administration-Finance/Payments/s4vu-giwb/data).

The raw dataset has ﻿746,106 rows and seven variables:

* VOUCHER NUMBER is regarded as the key of cases (although contains missing values)
* AMOUNT is numerical, describes the payment amount of each transaction (case)
* CHECK DATE is ordinal, describes when the payment is made
* Nominal variables are VENDOR NAME, CONTRACT NUMBER and CASHED (binary)

**Issues of the dataset**

From data analysis point of view

1. The number of variables that can be used in model building is very few. Only one numerical variable, the rest are nominal including two “keys”. Therefore, it becomes hard to calculate correlation matrix to view relationships among variables and gain insights for model building
2. Vender name and department name have too many levels to create dummy variables in the further analysis. I recoded department names into ordinal variables
3. 1964 rows of data have a negative number in AMOUNT column can slightly distort the overall distribution and violate data integrity
4. For rows labeled with “DV” in their contract number, the transaction is indirect with the payees. These hides feature and restricts pattern mining of indirect payees
5. The numbers in AMOUNT column have thousands separator. I encountered errors trying to convert the data type. I used Excel to modify data type on this column

From database point of view:

1. There are inconsistencies in recording of department names. 85% of values in this column is missing. After some initial data exploration, it’s discovered that there’s a partial dependency between missing department name and contract number. If contract number is coded “DV”, it’s likely that value for department names is null. I filled in these missing values with “DVC” (Direct Voucher Contract) to smooth the data.
2. The missing values in voucher number violate data integrity since it partially serves as a key to the dataset.  About 21% of values are missing in voucher number. Since this variable won't be included in the later stage, I did not take remedy actions.
3. Redundancy. There is also a partial redundancy between contract number of vendor name. Some payments were made under the same contract with a vendor, but each transaction was recorded in the dataset. Because there isn’t new information generated, cases under same contract name can be consolidated to one payment record. And details of each transaction can be accessed by making contract number as the foreign key.

**Techniques and Future Work**

I made use of Python packages such as Pandas, Numpy, sikit-learn and matlib.plot to analyze the dataset. By using Python group by function and aggregation functions I can view the results of payment amount variations among departments and payees; the payment amount fluctuations throughout this decade a list of long-term vendors and their transaction histories with the city and so forth. I used clustering techniques in order to find underlying payment records. The negative records in payment amount opens up the opportunity to construct an anomaly detection algorithm for the dataset by applying classifier techniques. Due to limited number of useful variables, I created few new categorical binary variables to facilitate the classifier. However, the fact that only one numerical variable in the model has restricted the clustering robustness. The performance matrices of both models are not as reliable.

The mining of the dataset can be more integral going forward. I would incorporate datasets from different data sources (datasets of government expending via other methods) to widen the variable choices so that a model can be developed to evaluate budget performance. In addition, it’ll improve the variable quality and reduce running time and complexity if the information of which industries vendors are in is included. And the outcomes of this model can help to check budget money allocation and reconstruct vendor pool. For instance, the amount of money paid to large corporations versus small to medium local businesses. The clustering model can be trained to find transactions related to the same project. Additionally, more time can be devoted to improving data visualization deliverables. For instance, an API that allows views to interact with data ranges so that the results of the analysis can be more transparent and interpretable to the public.

**Analysis, Results and Visualization**

1. Distribution of amount

Amount distribution is extremely skewed to right. 75% of the observations have amount less than $2497. 5% of the observations have amount over ﻿$145, 314. The most expensive money payments made directly to a vendor is to Blue Cross Blue Shield. Other expensive payments were made through a third party to indirect payees.



1. Fluctuations in total amount paid by year

The growth trend of average amount paid is not consist with that of the total amount paid.

From 2014 to 2018, the amount paid increased at a higher rate comparing to rates of 2010 to 2014. The fall of 2014 is probably infludence by 2014 mayoral election.

 

1. Top 10 payers



Throughout the spand of data was collected, top 10 departments that paid out the most are shown in the graph. The information can be used to evaluate budget plans.

1. Top 10 vendors



The graph shows the top 10 vendors of the city. Contraction companies are the ones that government does businesses the most.

1. Payment received by Blue Cross and Blue Shield



1. Amount paid out by O’Hare Modernalization Project



﻿