

Employee Indebtedness to the City of Chicago



ILLINOIS TECH

Utsav Pathak, Dhruv Dasadia, Dharmik Dharmesh Patel

College of Computing

Illinois Institute of Technology

30th November 2023

CS520

Data Integration, Warehousing, and Provenance

Dr. Boris Glavic

Fall 2023

Table of Contents

Abstract

1) Introduction.....	5
2) Relevance of the Dataset.....	6
a) Necessity of Debt Standardization	
b) Calculation and Repayment Mechanisms	
3) Attributes and Schema.....	7
a) Dataset Fields Overview	
b) Metadata Overview	
4) Discrepancies in the Dataset.....	9
a) The issue with Column Names	
b) Issue of missing values	
c) Possible miscalculation of debt	
5) Cleaning and Mechanisms.....	11
a) Understanding properties of every department	
b) Classifying ARMS ID	
c) Reassuring Removal Standards	
d) Removal of Unnecessary Rows	
6) Repopulation of Columns.....	14
a) Finding and Calculating New Employee Counts	
b) Cross verification of Changes	
7) ETL on the Dataset.....	15
8) Provenance.....	18
9) Limitations of the Analysis.....	18

10) Conclusion.....	19
11) Future Scope of Work.....	19

Abstract

The employee indebtedness to the city of chicago dataset has information on the public employees, categorized by the department, who owe money to the city as well as categorized departments and percentage of net debt owed.

In this project we take this data, clean and pre-process it to level the fields and debt mechanism and try to build an ETL process over it with respect to the federal debt data and try to understand the debt mechanism and atrocities in the data related to debt calculation.

1. Introduction

The Employee Indebtedness to the City of Chicago dataset is a public dataset that provides information on the number of City of Chicago employees who owe money to the City, as well as the total amount of debt owed by those employees. The dataset is updated weekly and includes data for all City of Chicago departments and sister agencies. The dataset includes the following information for each department or agency: Total number of individuals employed Number of individuals that owe funds to the City Total amount of debt owed by those employees The dataset can be used to analyze employee indebtedness trends over time, as well as to identify departments or agencies with high levels of employee indebtedness. The dataset can also be used to develop policies and programs to help City of Chicago employees reduce their debt.

The dataset that we consider for this project is from the mid of October 2011 to 09th November 2023 so that our cleaning mechanism does not need updating and we can cross validate if the mechanism is sufficient for the data that is updated later.

This is a non federal dataset which also means in case of debt, the dataset has higher number of issues with the standardization and in cases where the debt is soluble and paid to the federal authority directly or passes a clearance under from the federal net debt authority might have a higher chance of not getting updated, which therefore results in a mismatch between the data obtained from the chicago dataset website as well as the federal debt disclosure.

For the ETL Process, we take a look at the federal (historical debt outstanding to create a benchmark of debt) that can be found at [Link to Historical Debt Outstanding](#).

While ETL often involves merging, trying it with a dataset such as these comes with an additional issue such as change in debt calculation criterion and therefore, the datasets needs to be analyzed deeper to understand the debt calculation mechanism and both are to be brought on the same level of understanding.

While it does not look like a computational problem at the start, it actually is as it creates a necessity to have a pipeline to feed and clean data and bring it to the same level of normalization so that it gets easier to understand and better for decision making.

In this project, we try to achieve the initial stage of cleanup standards by pointing out the discrepancies and working on them to create a framework good enough to perform basic operations, which would just be the start to the entire pipeline of the larger mentioned problem of debt.

This can be cross validated by the fact that the dataset remains untouched in terms of work done on it on Kaggle, making it even more complex to find a start.

2. Relevance of the Dataset

a) Necessity of Debt Standardization

The dataset is an invaluable tool for academics, analysts, and policymakers since it offers a thorough picture of financial commitments made within the city. But the effectiveness of the dataset depends on how reliable and consistent the underlying data is, especially when it comes to debt computations. Standardizing debt calculation methods is essential to improving the dataset's dependability.

Large volumes of data must be stored and managed during the data warehousing process. Variations in debt calculation techniques might cause major distortions when analyzing the Chicago Indebtedness Dataset. Uniformity is ensured via standardization, making in depth analysis possible. This is especially important for evaluating the financial standing of various dataset organizations, including public institutions or local governments. Standardized methods for calculating debt also make data cleansing procedures easier. Cleaning is going over the dataset to find and fix any mistakes, outliers, or anomalies. If there are no uniform rules for calculations, cleaning turns into a laborious process that requires close examination of every entry to see if differences are indeed the consequence of real discrepancies or incorrect data entry. This procedure is made simpler by standardization, which also improves the dataset's general quality and expedites data cleaning activities.

The consequences for decision-making highlight the need for standardization. Therefore clean data is essential for policymakers to develop solutions that effectively handle economic concerns. Erroneous debt estimates may result in poor policy choices, which could exacerbate rather than resolve financial problems.

b) Calculation and Repayment Mechanisms

The way how debt is calculated in the dataset sets a lot of flags for each department as the dept for all the departments seems to be consistent but that for CPD (Chicago Police Department) is different, even with higher net employee and debt ratio, the net debt is less, this needs for accounting in factors of debt. The data provided should have a background on debt calculation mechanisms and if the debt is paid, how the revenue is made and if so, why not?

This helps in modeling perspectives of debt and to know which department is more liable to debt and which is not and which of them has specifically high debt. This would also help in understanding which of the public departments are more interactive towards the chicago public and which department in particular adds to the net federal debt.


More analysis would help break the entire dataset and benchmark smaller groups or bags of ARMS IDs towards one single ARMS ID. This can also help in the estimation of net debt ratio and employee count for that particular department.

3. Attributes and Schema

a) Dataset Fields Overview

The dataset has 7 attributes, Date, Department/Agency Name, ARMS Department ID, Total Number of Employees, Number of Employees with Debt, Percentage of Employees with Debt and Total amount Due.

Since the dataset is updated weekly, the number of rows increases, as of 09th November 2023, the dataset has 20593 rows.

employeeebtchicago (20593 rows) 

	Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)	Total_of_Employees (short)	_of_Employees_with_Debt (short)	Employees_with
0	06/04/22	ADMINISTRATIVE HEARING	AHMS	35	0	0
1	06/04/22	COMM ANIMAL CARE AND CONTROL	ANIMAL	61	1	1.6000000238418
2	06/04/22	AVIATION	AVIATION	?	39	2.0999999046325
3	06/04/22	BUS AFFAIRS AND CONSUMER PROT	BACP	168	2	1.2000000476837
4	06/04/22	BUILDINGS	BUILDINGS	233	2	0.8999999761581
5	06/04/22	CULTURAL AFFAIRS	CA	60	0	0
6	06/04/22	CHICAGO BOARD OF EDUCATION	CBOE	?	?	7
7	06/04/22	CITY COLLEGES OF CHICAGO	CCC	?	237	5.3000001907348
8	06/04/22	CCPSA	CCPSA	1	0	0
9	06/04/22	FIRE DEPARTMENT	CFD	?	96	2
10	06/04/22	CHICAGO HOUSING AUTHORITY	CHA	526	6	1.1000000238418
11	06/04/22	CITY CLERK	CLERK	78	2	2.5999999046325
12	06/04/22	DEPARTMENT OF PLANNING AND DEV	COMM_DEVEL	147	1	0.6999999880790
13	06/04/22	CIVILIAN OFFICE OF POLICE ACCOUNTABILITY	COPA	121	2	1.7000000476837
14	06/04/22	CITY COUNCIL	COUNCIL	351	18	5.0999999046325
15	06/04/22	CHICAGO PARK DISTRICT	CPDT	?	105	4.0999999046325
16	06/04/22	CHICAGO PUBLIC LIBRARY	CPL	?	4	0.4000000059604
17	06/04/22	CHICAGO TRANSIT AUTHORITY	CTA	?	?	23.200000762939
18	06/04/22	DAIS	DAIS	948	7	0.6999999880790
19	06/04/22	BOARD OF ELECTION COMMISSIONER	ELECTIONS	108	3	2.7999999523162

The dataset has several fields with NULL or No Values, these are mostly the ARMS_Department_ID, Total Number of Employees (Total_of_Employees), Number of Employees with Debt (_of_Employees_with_Debt) and Percentage of Employees with Debt (_Employees_with_Debt).

The good thing is that the dates in the dataset are consistent and all of the date fields are filled in the same order and none of them are empty or NULL.

[2]

≡

SELECT * FROM employeeebtchicago
WHERE Date IS NULL OR Date = '';

🔍

Console ▾

Timing

Datasets ▾

Charts ▾

🔗

📄

temporary_dataset (0 rows) 📄

Views > 📄 📄 📄

Date (string)

Department_or_Agency_Name (string)

ARMS_Department_ID (string)

Total_of_Employees (short)

_of_Employees_with_Debt (short)

_Employees_with_Debt (ffc

b) Metadata Overview

Column Name	Description		Type	
Date			Date & Time	📅
Data Type	API Field Name			
Floating Timestamp	date			
Department or Agency Name			Plain Text	T
Data Type	API Field Name			
Text	department			
ARMS Department ID	The department's ID in the City's ARMS system, which manages delinquent receivables. This field was created in July 2013 and will be blank for older records.		Plain Text	T
Data Type	API Field Name			
Text	arms_department_id			
Total # of Employees			Number	#
Data Type	API Field Name			
Number	dept_size			
# of Employees with Debt			Number	#
Data Type	API Field Name			
Number	employees			
% Employees with Debt			Number	#
Data Type	API Field Name			
Number	_employees_w_debt			
Total Amount Due			Number	#
Data Type	API Field Name			
Number	due			

4. Discrepancies in the Data

Inconsistencies in data reporting: The dataset is updated weekly, but the data is not always reported consistently. For example, the total amount of debt owed by employees may vary from one week to the next, even if the number of employees who are indebted to the city remains the same. Lack of transparency: The dataset does not provide any information about the reasons why employees are indebted to the city. This makes it difficult to understand the underlying causes of employee indebtedness and to develop effective strategies to reduce it.

Lack of transparency: The dataset does not provide any information about the reasons why employees are indebted to the city. This makes it difficult to understand the underlying causes of employee indebtedness and to develop effective strategies to reduce it.

In addition to these general discrepancies, the dataset also contains some specific inconsistencies. For example, the dataset shows that the Chicago Police Department has the highest number of employees who are indebted to the city, but the total amount of debt owed by CPD employees is lower than the total amount of debt owed by employees of other departments. This suggests that there may be some discrepancies in the way that debt is calculated or reported for CPD employees.

a) Issues with Column Names

The column names given in the dataset are as below, but while parsing in Vizier DB, the names get messed up in the CSV File and therefore require editing.

Column Name		
Date		
Department or Agency Name		
ARMS Department ID		
Total # of Employees		
# of Employees with Debt	Total_of_Employees	_Employees_with_Debt
% Employees with Debt		
Total Amount Due	_of_Employees_with_Debt	Total_Amount_Due

To sort this issue, we rename the column names.

Rename Column

Dataset employeeedebtchicago ▼

Column _Employees_with_Debt ▼

New Column Name PercentDebtEmployee

Rename Column

Dataset employeeedebtchicago ▼

Column _of_Employees_with_Debt ▼

New Column Name NumDebtEmployees

b) Issue of Missing Values

Now, we find that a lot of columns in Total_of_Employees, PercentDebtEmployee and NumDebtEmployee are NULL or do not have any values at all, we work on these and try to remove columns where we do not have the number of employees with debt as well as the percentage of employees with debt.

Total_of_Employees (short)	NumDebtEmployees (short)	PercentDebtEmployee (float)	Total_Amount_Due (string)
39	2	5.099999904632568	368
68	3	4.400000095367432	2,564.00
?	54	4.300000190734863	19,092.00
112	5	4.5	1,068.00
6	0	0	0
50	1	2	400
283	4	1.399999976158142	861
183	7	3.799999952316284	3,208.00
?	?	5.800000190734863	9,62,067.00
467	25	5.400000095367432	6,137.00
?	422	6.599999904632568	2,29,971.00
?	30	2.9000000953674316	8,147.00
?	?	19	6,42,713.00
98	3	3.0999999046325684	613
?	50	2.700000047683716	16,822.00

In rows where we do not have total, but the other two, we use the formula to populate values in total number of employees.

Where,

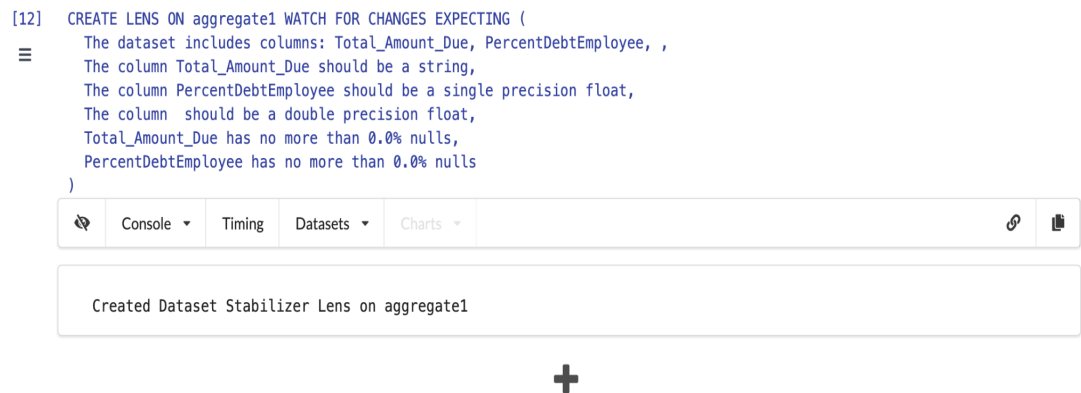
Total Number of Employees = (Percentage of Employees with Debt * 100)/Employees with Debt.

Similarly for Employees with Debt = (Percentage of Employees with Debt/100) * Total Number of Employees.

c) Possible Miscalculation of Debt

The dataset shows that the Chicago Police Department has the highest number of employees who are indebted to the city, but the total amount of debt owed by CPD employees is lower than the total amount of debt owed by employees of other departments. This suggests that there may be some discrepancies in the way that debt is calculated or reported for CPD employees.

To track and experiment with these, we created multiple lenses but weren't able to understand and differentiate between the debt calculation mechanism.



5. Cleaning and Mechanisms

Currently, we discover that many of the columns in Total_of_Employees, PercentDebtEmployee, and NumDebtEmployee are NULL or have no values at all. We attempt to eliminate them by identifying the columns that do not include the number of workers who are in debt or their percentage of debt. We apply the formula to populate values in the total number of workers in the two rows where we have total but not in the other one. The dataset reveals that while the number of CPD workers with outstanding municipal debt is more than that of employees from other departments combined, the total amount of debt owed by CPD employees is still less than that of the other departments combined.

a) Understanding properties of every department

Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)
06/04/22	ADMINISTRATIVE HEARING	AHMS
06/04/22	COMM ANIMAL CARE AND CONTROL	ANIMAL
06/04/22	AVIATION	AVIATION
06/04/22	BUS AFFAIRS AND CONSUMER PROT	BACP
06/04/22	BUILDINGS	BUILDINGS
06/04/22	CULTURAL AFFAIRS	CA
06/04/22	CHICAGO BOARD OF EDUCATION	CBOE
06/04/22	CITY COLLEGES OF CHICAGO	CCC
06/04/22	CCPSA	CCPSA
06/04/22	FIRE DEPARTMENT	CFD
06/04/22	CHICAGO HOUSING AUTHORITY	CHA
06/04/22	CITY CLERK	CLERK
06/04/22	DEPARTMENT OF PLANNING AND DEV	COMM_DEVEL
06/04/22	CIVILIAN OFFICE OF POLICE ACCOUNTABILITY	COPA
06/04/22	CITY COUNCIL	COUNCIL
06/04/22	CHICAGO PARK DISTRICT	CPDT
06/04/22	CHICAGO PUBLIC LIBRARY	CPL
06/04/22	CHICAGO TRANSIT AUTHORITY	CTA
06/04/22	DAIS	DAIS
06/04/22	BOARD OF ELECTION COMMISSIONER	ELECTIONS

	Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)
0	10/14/2011	Administrative Hearings	
1	10/14/2011	Animal Care & Control	
2	10/14/2011	Aviation	
3	10/14/2011	Board of Elections	
4	10/14/2011	Board of Ethics	
5	10/14/2011	Budget & Management	
6	10/14/2011	Buildings	
7	10/14/2011	Business Affairs & Consumer Protection	
8	10/14/2011	Chicago Board of Education	
9	10/14/2011	Chicago Housing Authority	
10	10/14/2011	Chicago Park District	
11	10/14/2011	Chicago Public Library	
12	10/14/2011	Chicago Transit Authority	
13	10/14/2011	City Clerk	
14	10/14/2011	City Colleges of Chicago	
15	10/14/2011	City Council	
16	10/14/2011	City Treasurer	
17	10/14/2011	Compliance	
18	10/14/2011	Cultural Affairs & Special Events	
19	10/14/2011	Emergency Management & Communications	

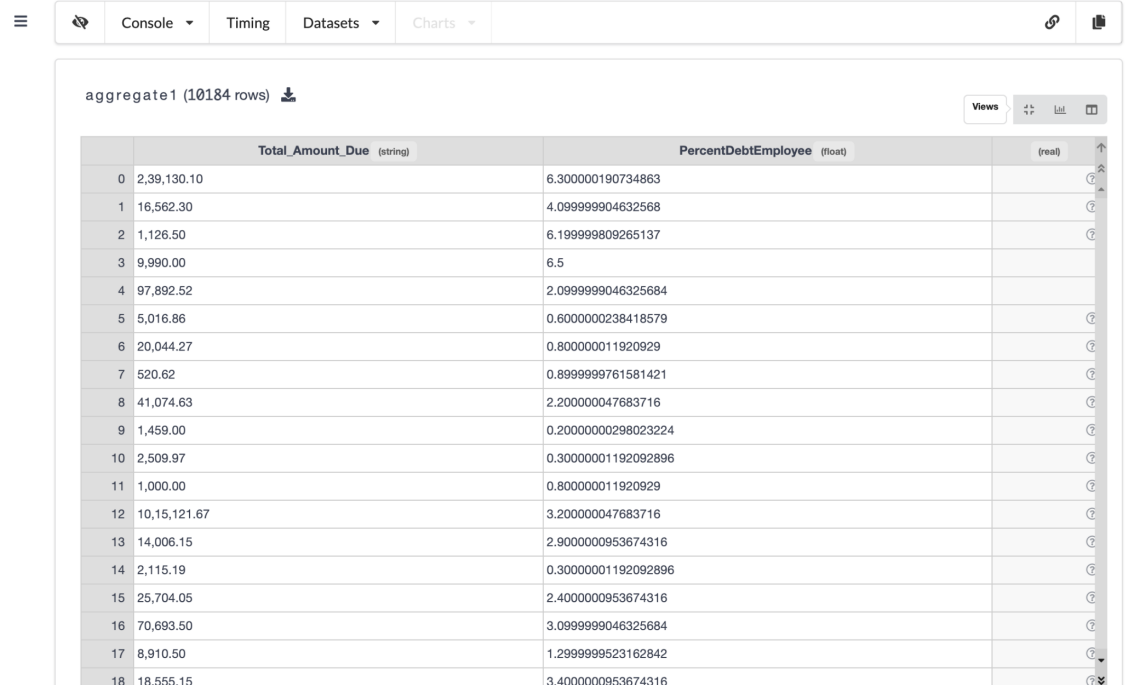
b) Classifying ARMS ID

We try to find ARMS IDs of the missing departments but this makes the provenance part too complex because of the creation of multiple duplicate entries of the same ID but different department name and therefore we dropped this idea.

c) Reassuring Removal Standards

We removed the rows where ARMS IDs and total employees or number of employees with debt and percentage were not present and also could not be calculated with the formula above or even with a median lens or the aggregation function.

```
[9] SELECT [6], [5], avg([2]) AS FROM employeeebtchicago GROUP BY [6], [5] INTO Aggregate1
```



	Total_Amount_Due (string)	PercentDebtEmployee (float)	(real)
0	2,39,130.10	6.300000190734863	
1	16,562.30	4.099999904632568	
2	1,126.50	6.199999809265137	
3	9,990.00	6.5	
4	97,892.52	2.0999999046325684	
5	5,016.86	0.6000000238418579	
6	20,044.27	0.800000011920929	
7	520.62	0.8999999761581421	
8	41,074.63	2.200000047683716	
9	1,459.00	0.20000000298023224	
10	2,509.97	0.30000001192092896	
11	1,000.00	0.800000011920929	
12	10,15,121.67	3.200000047683716	
13	14,006.15	2.9000000953674316	
14	2,115.19	0.30000001192092896	
15	25,704.05	2.4000000953674316	
16	70,693.50	3.0999999046325684	
17	8,910.50	1.2999999523162842	
18	18,555.15	3.4000000953674316	

d) Removal of Unnecessary Rows

We remove using the criteria as above from the table and obtain a new sheet of values.

6. Repopulation of Columns

a) Finding and Calculating new Employee Counts

We calculate and fill the new employee counts where the percentage of employees with debt and net employees with debt are present and fill in those values using the formula.

```
UPDATE employeeebtchicago
SET Total_of_Employees = (PercentDebtEmployee * 100)/NumDebtEmployees
WHERE Total_of_Employees IS NULL AND NumDebtEmployees IS NOT NULL AND PercentDebtEmployee IS NOT NULL;
```

b) Cross Verification of Changes

After removing and correcting values, we reprint the data.

	Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)	Total_of_Employees (short)	NumDebtEmployees (short)	PercentDebtEmployee (float)	
59	06/11/22	BOARD OF ETHICS	ETHICS	7	0	0	↑
60	06/11/22	FAMILY AND SUPPORT SERVICES	FAMILY_SUPPORT	560	9	1.600000023841858	↑
61	06/11/22	FINANCE	FINANCE	465	3	0.6000000238418579	↑
62	06/11/22	PUBLIC HEALTH	HEALTH	573	4	0.69999998079071	↑
63	06/11/22	HOUSING	HOUSING	78	1	1.2999999523162842	↑
64	06/11/22	COMMISSION ON HUMAN RELATIONS	HR	15	0	0	↑
65	06/11/22	HUMAN RESOURCES	HUMANRESOURCES	75	1	1.2999999523162842	↑
66	06/11/22	LAW	LAW	309	0	0	↑
67	06/11/22	OFFICE OF THE MAYOR	MAYOR	101	0	0	↑
68	06/11/22	MAYORS OFFICE-DISABILITIES	MAYOR_DISABIL	26	0	0	↑

[10]



Impute Missing Values

Dataset	employeeebtchicago		
Columns	Column	Model	
	Total_of_Employees	MultilayerPerceptron	
	PercentDebtEmployee	Median	
		<Pick One For Me>	



Change Command



Dismiss



Submit



Console

Timing

Datasets

Charts



Created Impute Missing Values Lens on employeeebtchicago

[11]

SELECT * FROM employeeebtchicago;

Console

Timing

Datasets

Charts

temporary_dataset (20593 rows)

Views

rtment_or_Agency_Name (string)	ARMS_Department_ID (string)	Total_of_Employees (short)	NumDebtEmployees (int)	PercentDebtEmployee (float)	Total_Amount_Due (string)
ID OF ETHICS	ETHICS	7	0	2.0999999046325684	0
Y AND SUPPORT SERVICES	FAMILY_SUPPORT	558	9	1.600000023841858	6,575.60
JCE	FINANCE	474	1	0.20000000298023224	70
IC HEALTH	HEALTH	562	2	0.4000000059604645	274.5
ING	HOUSING	77	1	1.2999999523162842	250
MISSION ON HUMAN RELATIONS	HR	15	0	2.0999999046325684	0
AN RESOURCES	HUMANRESOURCES	71	1	1.399999976158142	1,661.00
	LAW	301	0	2.0999999046325684	0
E OF THE MAYOR	MAYOR	103	1	1	240
RS OFFICE-DISABILITIES	MAYOR_DISABIL	27	0	2.0999999046325684	0

7. ETL Mechanisms on the Dataset

One of the most important steps in data management is extracting, converting, and loading (ETL) data into the desired database on the VizierDB once it has been retrieved from a variety of sources. We examine the ETL procedure in this instance to show how to glean valuable insights from unprocessed data.

Data extraction from the source is the initial stage of the ETL process. We read this data into a DataFrame and process it further to obtain more control over the data.

The data must next be transformed to make it appropriate for analysis after it has been extracted. This entails enhancing, organizing, and cleansing the data. For example, we may need to manage missing information, change the kind of data, or add additional functionality. Transformations in the Employee Indebtedness dataset involves controlling outliers, aggregating data, populating field and total values of employee and debt.

Importing the converted data is the last stage of ETL. This may be done using Python in a number of ways, such as uploading the DataFrame to a cloud storage service, saving it to a new CSV file, or putting it into an RDBMS. If the study calls for storage and saving data for future analytical work, one alternative is to load the data into the database and communicate with it using a database-specific library in SQL.

The abundance of Python's libraries becomes available for analysis and visualization once the data has been imported and converted. Explore data analysis with Pandas is possible, and meaningful visualizations may be produced with tools like Matplotlib and Seaborn. A custom script with scheduling capabilities or the use of technologies such as Apache Airflow to automate

the ETL process are two ways to guarantee that it stays efficient over time. By doing this, it is made sure that the analysis incorporates the most recent information and that it can comply with the weekly updates on the data.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = vizierdb.get_dataset('employeeebtchicago')
df['Date'] = pd.to_datetime(df['Date'])

df.dropna(subset=['Total_Amount_Due'], inplace=True)

total_amount_by_employee = df.groupby('ARMS_Department_ID')['Total_Amount_Due'].sum().reset_index()

total_amount_by_employee.to_csv('total_amount_by_employee.csv', index=False)

sns.barplot(x='ARMS_Department_ID', y='Total_Amount_Due', data=total_amount_by_employee)
plt.title('Total Indebtedness by Employee')
plt.xlabel('ARMS_Department_ID')
plt.ylabel('Total_Amount_Due')
```

	Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)
0	10/14/2011	Administrative Hearings	
1	10/14/2011	Animal Care & Control	
2	10/14/2011	Aviation	
3	10/14/2011	Board of Elections	
4	10/14/2011	Board of Ethics	
5	10/14/2011	Budget & Management	
6	10/14/2011	Buildings	
7	10/14/2011	Business Affairs & Consumer Protection	
8	10/14/2011	Chicago Board of Education	
9	10/14/2011	Chicago Housing Authority	
10	10/14/2011	Chicago Park District	
11	10/14/2011	Chicago Public Library	
12	10/14/2011	Chicago Transit Authority	
13	10/14/2011	City Clerk	
14	10/14/2011	City Colleges of Chicago	
15	10/14/2011	City Council	
16	10/14/2011	City Treasurer	
17	10/14/2011	Compliance	
18	10/14/2011	Cultural Affairs & Special Events	
19	10/14/2011	Emergency Management & Communications	

Date (string)	Department_or_Agency_Name (string)	ARMS_Department_ID (string)
06/04/22	ADMINISTRATIVE HEARING	AHMS
06/04/22	COMM ANIMAL CARE AND CONTROL	ANIMAL
06/04/22	AVIATION	AVIATION
06/04/22	BUS AFFAIRS AND CONSUMER PROT	BACP
06/04/22	BUILDINGS	BUILDINGS
06/04/22	CULTURAL AFFAIRS	CA
06/04/22	CHICAGO BOARD OF EDUCATION	CBOE
06/04/22	CITY COLLEGES OF CHICAGO	CCC
06/04/22	CCPSA	CCPSA
06/04/22	FIRE DEPARTMENT	CFD
06/04/22	CHICAGO HOUSING AUTHORITY	CHA
06/04/22	CITY CLERK	CLERK
06/04/22	DEPARTMENT OF PLANNING AND DEV	COMM_DEVEL
06/04/22	CIVILIAN OFFICE OF POLICE ACCOUNTABILITY	COPA
06/04/22	CITY COUNCIL	COUNCIL
06/04/22	CHICAGO PARK DISTRICT	CPDT
06/04/22	CHICAGO PUBLIC LIBRARY	CPL
06/04/22	CHICAGO TRANSIT AUTHORITY	CTA
06/04/22	DAIS	DAIS
06/04/22	BOARD OF ELECTION COMMISSIONER	ELECTIONS

[7]
≡

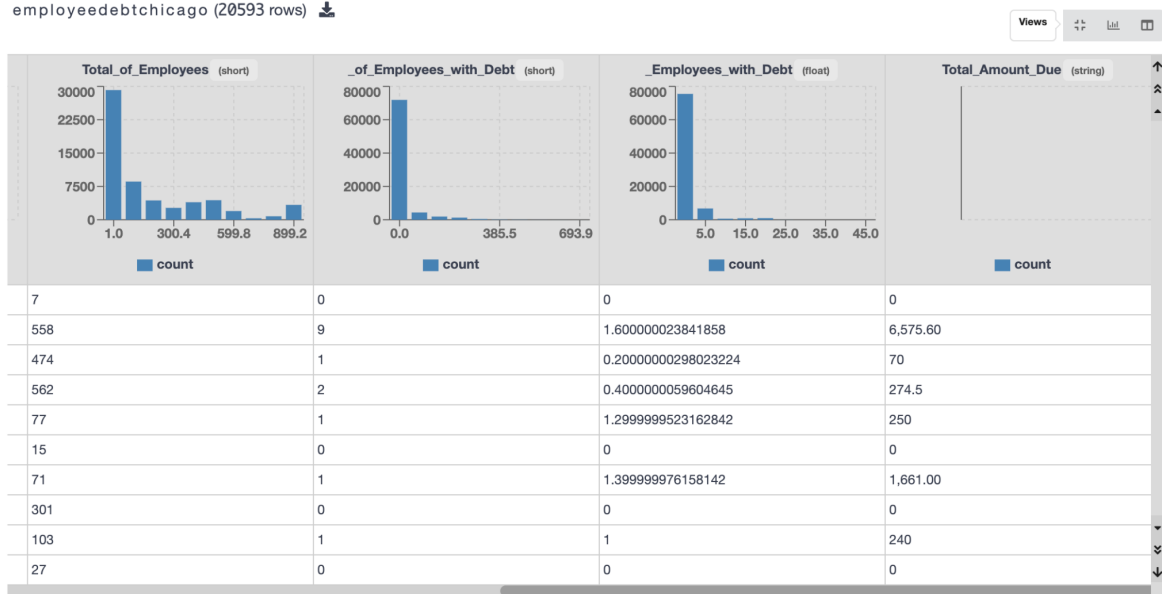
```
SELECT Department_or_Agency_Name, SUM(Total_Amount_Due) as total_debt
FROM employeedebtchicago GROUP BY Department_or_Agency_Name;
```

Console Timing Datasets Charts

temporary_dataset (100 rows)

	Department_or_Agency_Name (string)	total_debt (real)
0	BUS AFFAIRS AND CONSUMER PROT	63625.170000000001
1	HOUSING	21374.600000000013
2	Family & Support Services	4377.03
3	Fire Department	
4	Administrative Hearings	4484.65
5	COMMISSION ON HUMAN RELATIONS	6532.400000000001
6	Chicago Public Library	3253.7799999999997
7	Office of the Mayor	4956.44
8	FIRE DEPARTMENT	
9	Library	1580.0200000000002
10	OFFICE OF EMERGENCY MGMT & COM	
11	Streets and Sanitation	
12	Board of Ethics	296.91999999999996
13	Mayor's Office for People with Disabilities	1961.2
14	CHICAGO BOARD OF EDUCATION	
15	Chicago Transit Authority	
16	Administrative Hearings	1333.03

employeeebtchicago (20593 rows) 



8. Provenance

Provenance (in case of time and date) is really good with this data, although the major issue arises when we try to find origin of the information, calculation metrics and the quantity, this is ensured with the anomalies in debt calculation of CPD.

The metadata is clear and so are the dependencies that helped us in populating the columns of the data such as total number of employees and number of employees with debt, this aided as we prevented almost half of the NULL Fields from being dropped out of the table.

We have no idea about the interactions and update on the data, that is if debt is removed or not, from what we obtain, is that new debt is added but old debt is not removed or their is no way to trace of it is paid or vacated by federal reserves, which is a major issue in understanding if the net debt obtained is true or not.

9. Limitations of the Analysis

Our analysis limits to handling and filling the employees count and creating an ETL with the federal debt sheet.

We also try to come up with princesses to help calculate and mark the debt better and focus on specific departments such as the CPD and als try to find the missing ARMS IDs

and try and check if they map with any existing ARMS IDs or are sub departments of a larger department.

10. Conclusion

We realized that a lot less work has been done on the debt data, at least for the part which is publicly available and there is no background on how the data is collected or kept and maintained, which causes a lot of issues and anomalies that are expressed earlier in the report.

This issue caused the larger part of this project to be spent on understanding and cleaning the data, which we only got to know as we started working on the data further and had not expected the anomalies to be so large at the beginning of the project.

11. Scope of Future Work

We need a better understanding of the dataset and how the debt is calculated and reported, which will further help in understanding the data and help create a proper analytics dashboard.

The dataset is updated weekly and some new issues may arise in the data, so we would need to create a mechanism to understand these new changes and create new filters and a fault tolerant mechanism for the analytical dashboard (if created) to be responsive.
