UNIVERSITY OF ILLINOIS
URBANA-CHAMPAIGN

CS513 Theory and Practice of Data Cleaning

Summer 2023

Final Project: Phase-1 Report

Team 19
Hermann Rösch (NET ID: hrosch2@illinois.edu)
Lakshmi Susheela Amrutha Pydeti (NET ID: lpydeti2@illinois.edu)

§ 1. Dataset Chosen:

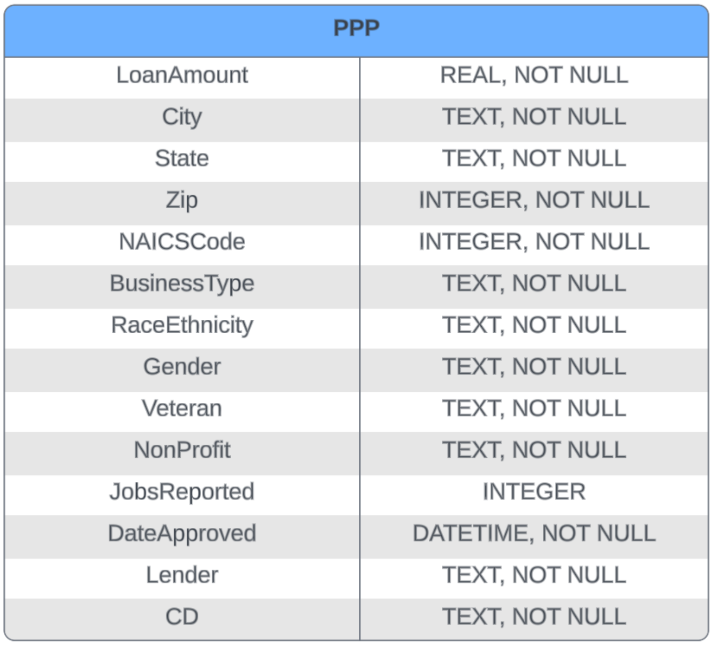
We have selected the **Paycheck Protection Program (PPP) dataset** for our project.

§ 2. Description of Dataset:

The PPP dataset is a comprehensive record encapsulating information about loans under the Paycheck Protection Program (PPP) provided by commercial lending banks to small businesses backed by the Federal government. Our current dataset contains only information related to the state of Hawaii. The PPP initiative was the U.S. government's response to the economic distress caused by the COVID-19 pandemic. The dataset incorporates business, loan, and demographic information, reflecting business operations and trends across various sectors in Hawaii.

- **Temporal Extent:** The dataset contains data about the Paycheck Protection Program, which began in April 2020 and ended in May 2021. However, the dataset provided for analysis only contains data from April 2020 to August 2020 and does not include any information about the businesses or their financial situations before or after these dates.
- **Spatial Extent:** The dataset is limited to businesses in Hawaii. It does not include businesses from other states, U.S. territories, or foreign countries.
- **Origin of Data:** The data is gathered from the loan applications submitted to the Paycheck Protection Program. The data is self-reported by the businesses and collected and published by the Small Business Administration (SBA), a U.S. government agency. The dataset is sourced from the Investigative Reporting Workshop.

The dataset's structure can be represented as a relational database schema, with each row corresponding to a unique loan application and each column representing a distinct attribute of the application. Below is a picture that illustrates the PPP dataset as a table and its column data types:



PPP	
LoanAmount	REAL, NOT NULL
City	TEXT, NOT NULL
State	TEXT, NOT NULL
Zip	INTEGER, NOT NULL
NAICSCode	INTEGER, NOT NULL
BusinessType	TEXT, NOT NULL
RaceEthnicity	TEXT, NOT NULL
Gender	TEXT, NOT NULL
Veteran	TEXT, NOT NULL
NonProfit	TEXT, NOT NULL
JobsReported	INTEGER
DateApproved	DATETIME, NOT NULL
Lender	TEXT, NOT NULL
CD	TEXT, NOT NULL

Below is a short description of each of the columns in the dataset:

Column	Description
LoanAmount	The loan amount granted under the PPP program is represented as a float number
City	The city where the business that received the loan is located
State	The state where the business is located. In this dataset, the state is always Hawaii (HI)
Zip	The ZIP code where the business is located in Hawaii
NAICSCode	The North American Industry Classification System code for the business. It represents the sector in which the business operates
BusinessType	The type or structure of the business, such as "Corporation", "Limited Liability Company", "Non-Profit Organization", and others
RaceEthnicity	The self-reported race or ethnicity of the business owner. Possible values include specific races, "Unanswered", and others
Gender	The gender of the business owner. It can be "Male Owned", "Female Owned" or "Unanswered"
Veteran	Indicates the veteran status of the business owner. Possible values are "Veteran", "Non-Veteran", or "Unanswered"
NonProfit	Indicates whether the business is a non-profit organization. This field will contain a 'Y' if the business is non-profit; if not, it will be left blank
JobsReported	The number of jobs reported to be retained due to the loan
DateApproved	The date when the loan was approved, in the format MM/DD/YYYY
Lender	The name of the bank or financial institution that approved the loan
CD	The Congressional District is where the business is located. The code is in the format 'HI-XX', where 'XX' is the district number

§ 3. Use Cases:

- **"Zero data cleaning" use case (U0):** Without any cleaning, we can answer the question, **"What is the total amount of loans provided?"** This question can be answered directly from the `LoanAmount` column without cleaning or manipulating any other data in columns.

- **"Main" use case (U1):** Data cleaning becomes necessary for complex analysis like **"Identify the cities that received the highest total amount in PPP loans."** This use case is particularly interesting because it highlights which cities in Hawaii had a higher concentration of businesses applying for loans and could indicate areas that were most impacted by the pandemic. However, achieving this requires a comprehensive review and cleaning of city names and ZIP codes to ensure accurate aggregation of loan amounts.

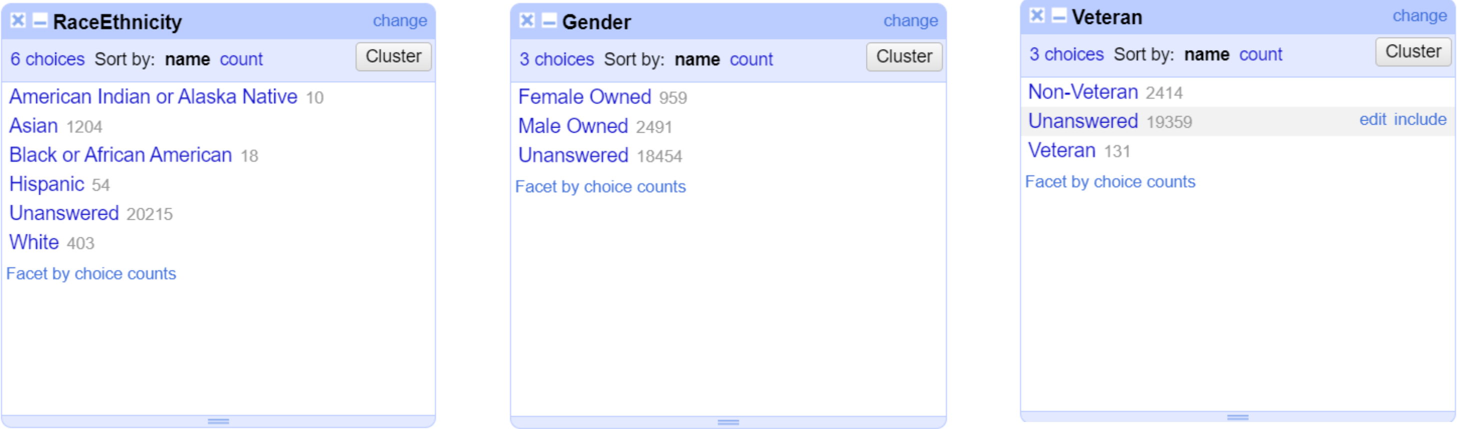
- **"Never enough" use case (U2):** Even with extensive cleaning, we cannot answer the question **"What is the repayment status of businesses?"** This question would require additional data not present in the current dataset, such as loan repayment records or financial statements from the businesses.

§ 4. Data Quality Problems

Several data quality problems are evident from a preliminary inspection of the dataset:

Missing Values:

- The `RaceEthnicity`, `Gender`, and `Veteran` columns have numerous instances of the value **"Unanswered"**. This may indicate that the respondents chose not to provide this information; however, from the perspective of data analysis, these are treated as missing values that limit the demographic analysis that can be performed.



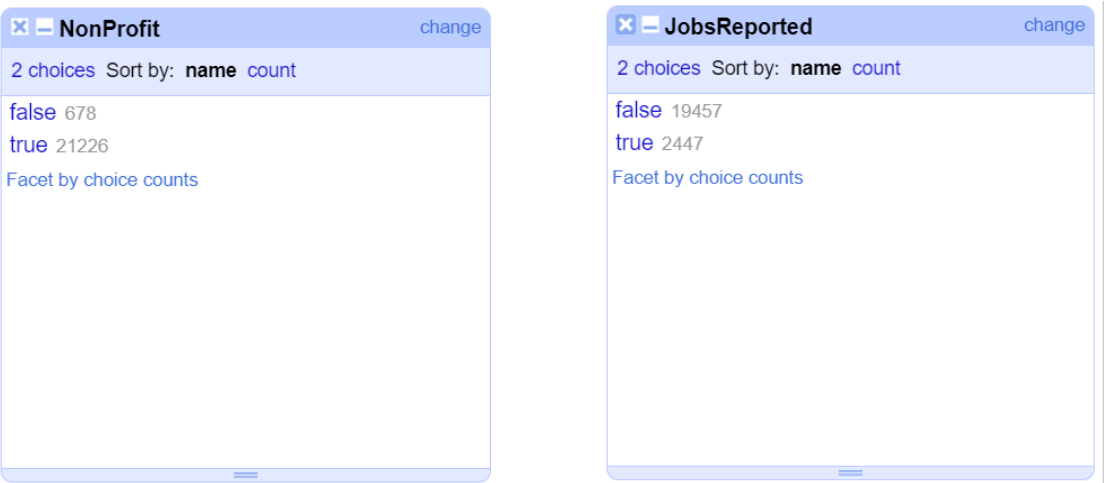
RaceEthnicity
6 choices Sort by: name count
American Indian or Alaska Native 10
Asian 1204
Black or African American 18
Hispanic 54
Unanswered 20215
White 403
Facet by choice counts

Gender
3 choices Sort by: name count
Female Owned 959
Male Owned 2491
Unanswered 18454
Facet by choice counts

Veteran
3 choices Sort by: name count
Non-Veteran 2414
Unanswered 19359
Veteran 131
Facet by choice counts

Figure 1 - Unanswered values in `RaceEthnicity`, `Gender` and `Veteran` columns

- On the other hand, the `NonProfit` and `JobsReported` columns contain blank values representing missing data that can affect the operational analysis of businesses.



NonProfit
2 choices Sort by: name count
false 678
true 21226
Facet by choice counts

JobsReported
2 choices Sort by: name count
false 19457
true 2447
Facet by choice counts

Figure 2 - Missing (blank) values in `NonProfit` and `JobsReported` columns

Incorrect Data: Some `City` names are associated with wrong `Zip` codes, leading to inaccurate geographical analysis, as `Zip` codes are a key indicator of a business's location. Verifying the accuracy of `Zip` codes against a reliable city-to-ZIP code mapping library is suggested to address this problem.

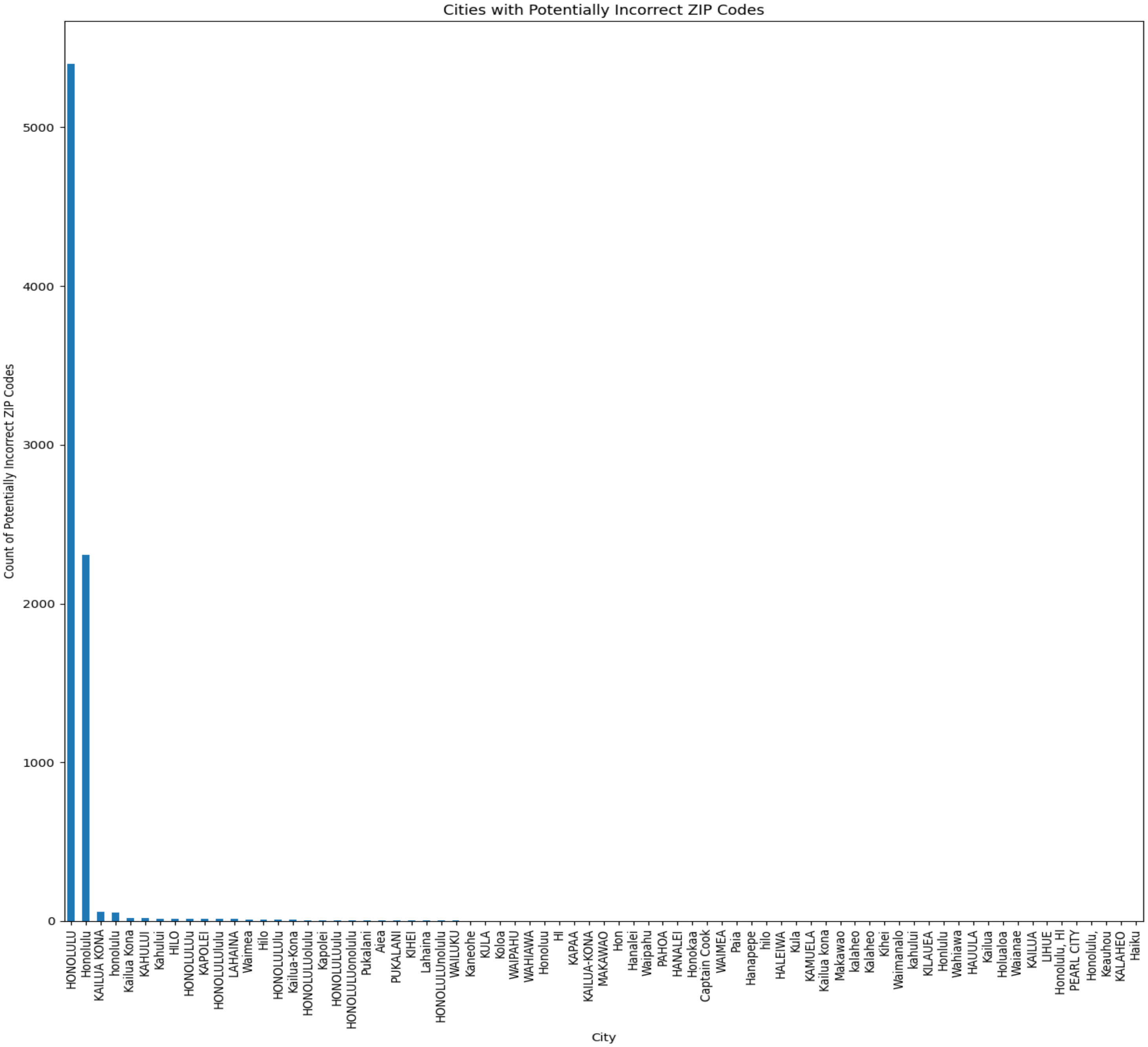
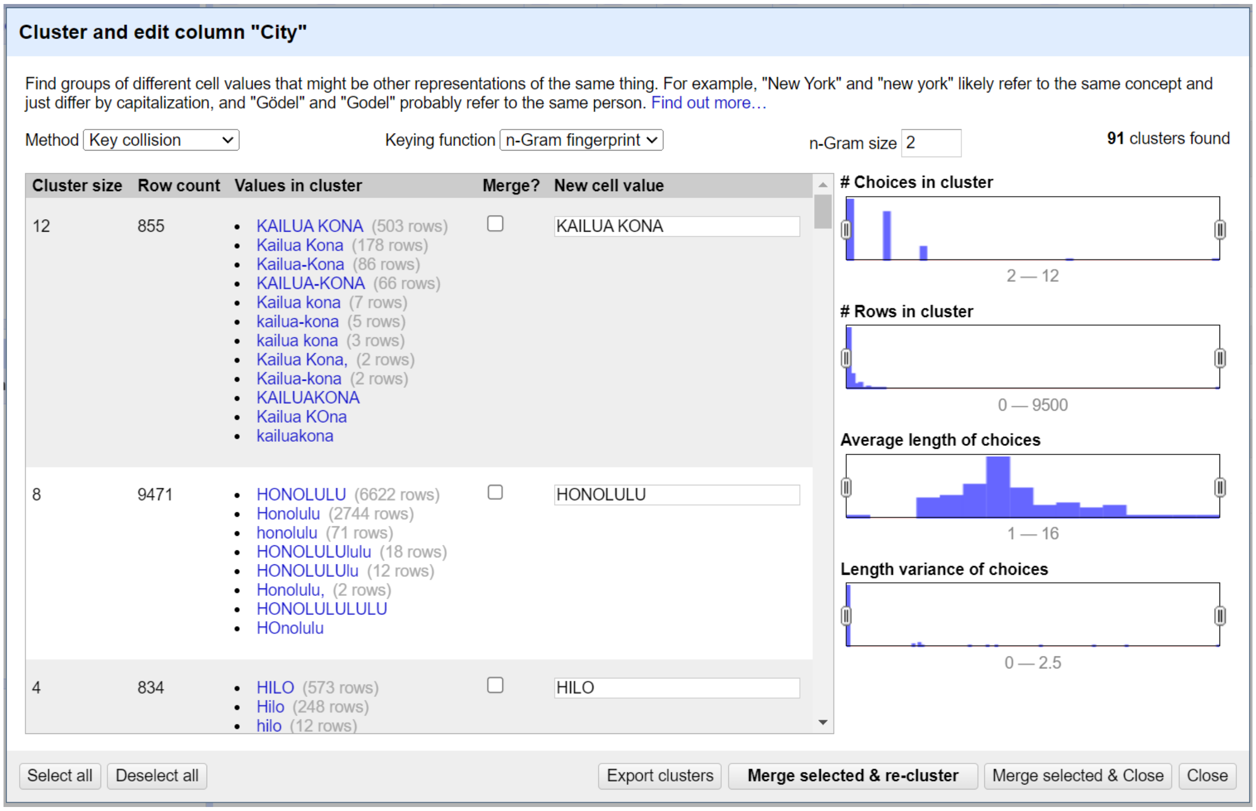


Figure 3 - Cities associated with potentially incorrect Zip Codes

Inconsistencies: The `City` column contains inconsistent entries due to variations in capitalization and spelling. For accurate geographical analysis, these inconsistencies need to be standardized.



Cluster and edit column "City"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: Key collision — Keying function: n-Gram fingerprint — n-Gram size: 2 — **91** clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
12	855	KAILUA KONA (503 rows) Kailua Kona (178 rows) Kailua-Kona (86 rows) KAILUA-KONA (66 rows) Kailua kona (7 rows) kailua-kona (5 rows) kailua kona (3 rows) Kailua Kona, (2 rows) Kailua-kona (2 rows) KAILUAKONA Kailua KOna kailuakona	☐	KAILUA KONA
8	9471	HONOLULU (6622 rows) Honolulu (2744 rows) honolulu (71 rows) HONOLULUlulu (18 rows) HONOLULUlu (12 rows) Honolulu, (2 rows) HONOLULULULU HOnolulu	☐	HONOLULU
4	834	HILO (573 rows) Hilo (248 rows) hilo (12 rows)	☐	HILO

Figure 4 - Variations in capitalization and spelling in `City` column values

Why is data cleaning necessary to support the "Main" use case U1?

Data cleaning is crucial for the **"Main" use case (U1)**. Since we aim to **"Identify the cities that received the highest total amount in PPP loans,"** inconsistencies in city names and incorrect ZIP codes could severely impact the accuracy of our results. By cleaning these attributes, we can aggregate loan amounts accurately per city. Without this step, we might erroneously report the total loan amounts received by a city, undercounting or overcounting because of spelling inconsistencies or incorrect ZIP codes.

§ 5. Initial Plan for Phase-II

S1 Review and update Use Case and Dataset Description

Both team members will revisit and revise the use case and the dataset description as necessary to ensure that it captures the current status and understanding of the dataset.

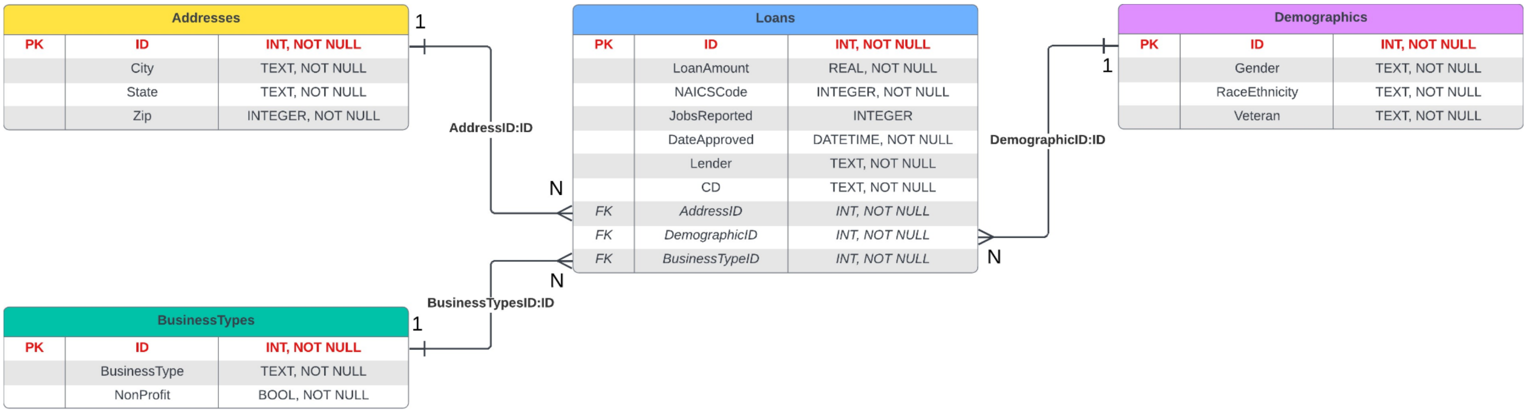
The PPP dataset contains multiple columns with information about the businesses, demographic details, loan amounts, and other relevant attributes. However, considering the PPP dataset as a single table introduced complexities due to the heterogeneous nature of the information and the presence of redundant and missing data.

After following all the data cleaning steps (listed in the upcoming steps), we plan to split the original PPP dataset into four (4) tables: `Addresses`, `BusinessTypes`, `Loans`, and `Demographics` to enhance data manageability and improve query performance. Apart from that, the decision to split the dataset also was driven by the distinct categories of information these tables represent:

- `Addresses`: Provide location-specific details, crucial for accurately aggregating loan amounts by City.
- `BusinessTypes`: Isolates the type of each business, potentially useful for the analysis of loan distributions across different business categories
- `Demographics`: Encapsulates demographic details associated with loans, providing the potential for socio-economic analysis or impact studies.
- `Loans`: The **"main table"** that contains loan-specific information, functioning as a central hub connecting to all other tables. It provides a clearer view of the financial aspects intrinsic to the PPP dataset.

Splitting the data into these four tables reduces redundancy and promotes scalability for future changes or additions. Furthermore, it allows for enforcing data integrity constraints by defining relationships, primary keys, and foreign keys across tables, ensuring data consistency and accuracy.

Below is the ER Diagram that represents the relationships between the `Addresses`, `BusinessTypes`, `Demographics,` and `Loans` tables:



	Addresses	
PK	ID	INT, NOT NULL
	City	TEXT, NOT NULL
	State	TEXT, NOT NULL
	Zip	INTEGER, NOT NULL

	Loans	
PK	ID	INT, NOT NULL
	LoanAmount	REAL, NOT NULL
	NAICSCode	INTEGER, NOT NULL
	JobsReported	INTEGER
	DateApproved	DATETIME, NOT NULL
	Lender	TEXT, NOT NULL
	CD	TEXT, NOT NULL
FK	*AddressID*	*INT, NOT NULL*
FK	*DemographicID*	*INT, NOT NULL*
FK	*BusinessTypeID*	*INT, NOT NULL*

	Demographics	
PK	ID	INT, NOT NULL
	Gender	TEXT, NOT NULL
	RaceEthnicity	TEXT, NOT NULL
	Veteran	TEXT, NOT NULL

	BusinessTypes	
PK	ID	INT, NOT NULL
	BusinessType	TEXT, NOT NULL
	NonProfit	BOOL, NOT NULL

The above dataset restructuring steps are in line with our **"Main" use case (U1)**. We facilitate accurate aggregation and analysis by ensuring the accuracy of city names and ZIP codes in the `Addresses` table and linking loan amounts from the `Loans` table.

We aim to complete S1-Reviewing and Updating our Use Case and Dataset description by 7/12/2023.

S2 Identify DQ Problems using OpenRefine and Manual Inspections

Both team members will analyze the original dataset to identify these DQ problems. The initial PPP dataset has data quality and inconsistency issues that may affect data curation. Some of them are listed below:

- Unnecessary white spaces are present in some columns, leading to potential inconsistencies in data values.

- There are missing (blank) values in `JobsReported` and `NonProfit` columns. *(Figure-2)*

- Inaccurate associations between `City` names and `Zip` codes exist in the dataset, possibly due to data entry errors or outdated information. *(Figure-3)*

- Inconsistencies in the `City` column, including typos, and variations in the casing (lowercase, uppercase), make identifying unique city names challenging and cause issues in detecting duplicate clusters *(Figure-4).* These inconsistencies need to be standardized for accurate geographical analysis.

- The date format for the column `DateApproved` is not in the standard ISO date format.

We will primarily use `OpenRefine`, an open-source application known for its data-cleaning capabilities. `OpenRefine` offers features like clustering and General Refine Expression Language (GREL), similar to regular expressions, which will prove essential for addressing the identified data quality issues. Once the file is clean enough, we export all the operations history to a `JSON` file.

However, given the dataset's complexity and various data quality issues, we will complement `OpenRefine` functionalities with `Python` libraries like `pandas` and `uszipcode` for more efficient and accurate data cleanup in subsequent steps.

We expect S2-Identifying potential Data Quality problems to be completed by 7/15/2023.

S3 Perform Data Cleaning

Amrutha will use `OpenRefine` and `Python pandas` and `uszipcode` libraries to perform bulk transformations and data-cleaning operations on the original PPP dataset. The cleaning process will include:

Trimming white spaces: Unnecessary white spaces in relevant columns will be removed using OpenRefine.

Fixing inconsistent cases: The case of the `City` column will be converted to uppercase using `OpenRefine` to address inconsistencies.

Correcting City names: `OpenRefine` clustering capabilities will be used to make a facet and perform the cluster operation using the key-collision method and fingerprint function. Then we will merge the relevant clusters, and repeat the process with n-gram fingerprint, metaphone3, and cologne-phonetic methods. Finally, we will make a facet and perform the cluster operation using nearest neighbor and Levenhstein distance function.

Date transformations: `OpenRefine` will be used to convert the date to standard ISO format in the DateApproved column.

NonProfit column transformation: Python's `pandas` library's **`fillna`** function will be used to replace missing values with **False** in the `NonProfit` column, while **'Y'** values will be replaced with **True** using the `replace` function.

Removing duplicate records: The `pandas` **`drop_duplicates`** method will be used to eliminate duplicate records from the dataset.

Improve `City` and `Zip` associations using the `uszipcode` library: The `uszipcode` library provides accurate mapping based on multiple data sources including the U.S. Census Bureau and USPS. We will use the `uszipcode` library to correct discrepancies in city and ZIP code associations.

We anticipate the completion of S3-Peform Data Cleaning by 7/18/2023.

S4 Data Quality Checking

Following the data cleaning process, Hermann will use Python's `pandas` and `uszipcode` libraries to develop test cases and illustrate the improvement in data quality; this will verify that the cleaned data properly adheres to the updated schema and integrity constraints. Below are some of the planned test cases:

Ensure correction in misspelled `City` entries: `OpenRefine` clustering algorithms will be employed to identify and merge inconsistent city names caused by variations in capitalization and spelling errors. The demonstration of a subset of standardized entries will illustrate the improvements in data quality.

Correction Verification in City entries: A subset of `City` entries will be selected from the cleaned dataset, and each entry will be checked for consistency in spelling and capitalization using Python string functions within pandas, illustrating the effectiveness of the `OpenRefine` cleaning process.

City and Zip associations check: After cleaning the entire PPP dataset, we'll validate the corrections on a subset of `Zip` entries against the accurate associations provided by the `uszipcode` library. This will validate the effectiveness of the city-ZIP code association correction process in the cleaned data.

NonProfit column check: The `value_counts` function in `pandas` will be used to confirm that the `NonProfit` column in the cleaned dataset contains only boolean values **(True/False)**, verifying the successful transformation of this column.

Data redundancy check: Using the `duplicated` function in `pandas`, we will confirm that there are no duplicate rows in the cleaned dataset, validating the successful reduction of data redundancy

"Main" use case (U1) support check: An aggregation query will be executed using `pandas groupby` and `sum` functions to calculate the total loan amount received by each city in both the original and cleaned datasets. Then, the `nlargest` function will be used to identify the top 5 cities that received the highest total loan amounts.

By comparing these two sets of results, we can validate the effectiveness of our data cleaning process, especially the accuracy improvements in city names and ZIP code associations. This serves as a practical test case for the real-world relevance and utility of our data-cleaning operations.

We aim to complete S4-Data Quality Checking by 7/20/2023.

S5 Document and Quantify Change

In the final step of the process, we will use the `YesWorkflow` tool to document the changes made during the data cleaning process comprehensively. `YesWorkflow's` ability to visualize workflows will be beneficial to accurately represent the flow of data transformations and cleaning operations performed in previous steps.

We will use `YesWorkflow` to create a diagram of the data transformation and cleaning steps. This diagram will help us to visualize and understand the flow of data from the raw PPP dataset to the final cleaned, transformed, and split/restructured dataset.

We will then quantify the changes by comparing the initial PPP dataset **(D)** and the cleaned and split/restructured dataset **(D')** regarding the number of rows, columns, and values within these columns before and after each process. This comparison will also include detecting integrity constraint violations before and after the data cleaning operations.

Through this documentation, we can create a comprehensive narrative of the cleaning process, documenting the types and amounts of changes that occurred from **D** to **D'**. This thorough documentation will be a valuable reference for future data analysis or audits, providing a clear audit trail of the changes made.

We expect S5-Document and Quantify Change to be completed by 7/22/2023.