CS513 Theory & Practice of Data Cleaning Final Project

Project Phase-I

Team members:
Xin Peng (xinp2@illinois.edu)

Amy Zhao (yiminz3@illinois.edu)

Dajun Lin (dajunl2@illinois.edu)

1. Identify a dataset:

We decided to use the PPP-data.

2. Develop a target (main) use case

- Main use case (data cleaning is necessary): Plot Loan Amount/Count by City.

  We would like to investigate the distribution of loan averages by City. This is calculated by average load amount. We would also like to know the total sum of the loan about by city. If a city has a higher average loan average, it means that for business in that city got larger loan per business. If a city has a larger total loan amount, it means that city received more loans than other cities and the business in that city have received more loans. Knowing the loan distribution can have a better view of the business situation that needs the loan.

- Minor use case 1 (good enough as is): Loan by “Zip” code

  For “Zip” column, there is no missing data. Moreover, “Zip” can be used to calculate the loan averages in a region, which has practical significance for the next step of data analysis.

- Minor use case 2 (never good enough): Loan by RaceEthnicity:

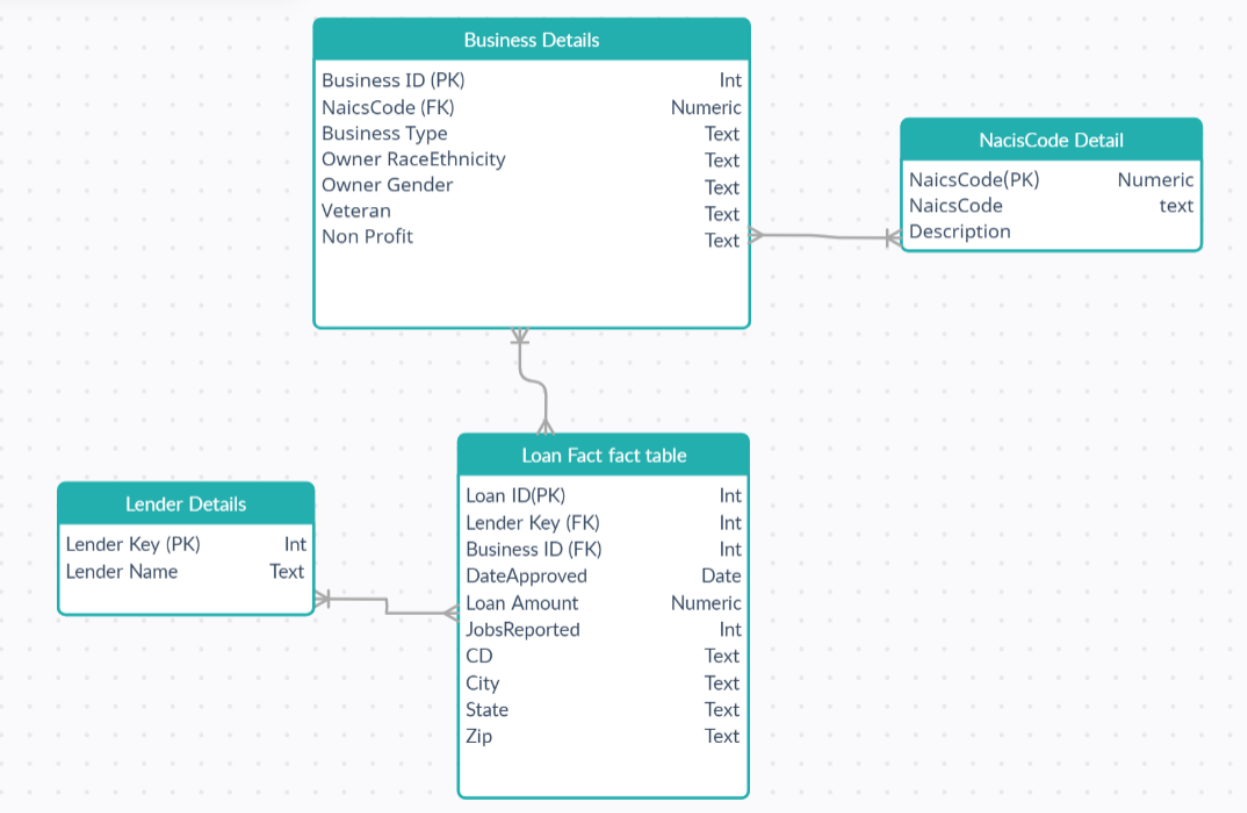
  This is a voluntary field. 90% of the data is “Unanswered”. Therefore, grouping by RaceEthinicity does not provide meaningful analysis.

3. Describe the dataset

The data is a list of PPP loans in the amount of up to $150K, approved from April to August 2020, in Hawaii State. PPP is an SBA-backed loan that helps businesses keep their workforce

employed during the COVID-19 crisis.

The data schema is drawn as below:



We identified 1 fact table and 3 dimension tables as follows:

Fact table with the following fields with datatypes:

Column Name	Data Type	Nullable	Description
Loan ID (PK)	Integer	Not Null	Primary Key. Unique identifier for loans.
Lender Key (FK)	Integer	Null	Foreign Key that joins to {Lender Detail} dimension table
Business ID (FK)	Integer	Null	Foreign Key that joins to {Business Detail} dimension table
DateApprovel	Date	Not Null	Loan approval date
Loan Amount	Numeric	Null	Loan amount
JobsReported	Integer	Null	Jobs reported in the loan application
CD	Text	Not Null	Loan code
City	Text	Not Null	City of the loan

State	Text	Not Null	State of the loan
Zip	Text	Not Null	Zip code of the loan

Three Dimension tables with separate tables are:

1. Business Details table:

Column Name	Data Type	Nullable	Description
Business ID (PK)	Integer	Not Null	Primary Key. Unique identifier for business
NaicsCode(FK)	Numeric	Null	Foreign Key that joins to {NaicsCode Detail} dimension table
Business Type	Text	Null	Business Type
Owner RaceEthinicity	Text	Not Null	Business owner's race or ethnicity. "Unanswered" if no answer entered
Owner Gender	Text	Not Null	"Female Owned", "Maile Owned", or "Unanswered"
Veteran	Text	Not Null	"Veteran", "Non-Veteran", or "Unanswered"
NonProfit	Text	Null	"Y" for non-profit business

2. Lender Detail table:

Column Name	Data Type	Nullable	Description
Lender key (PK)	Integer	Not Null	Primary Key. Unique identifier for Lender
Lender Name	Text	Not Null	Lender's name

3. NaicsCode Detail table:

Column Name	Data Type	Nullable	Description
NaicsCode (PK)	Numeric	Not Null	Primary Key. Unique identifier for business industry
NaicsCode Description	Text	Not Null	Industry description

4. List obvious data quality problems

We found these data columns with issues presented in the data:

- City column: not case consistent & incorrect value.
  E.g., "184 Puueo Street", "5", "_" , "1137 11th Avenue", can use zip code to find correct city

- Zip and City unmatched.
  E.g., zip code 96749 has city Hilo in the dataset but in fact it is not associated with Hilo.

- NaicsCode: missing values

5812	WAIPAHU	HI	96797	999990
5688	Waipahu	HI	96797	999990
57967	AIEA	HI	96701	
22915	AIEA	HI	96701	
15555	ANAHOLA	HI	96703	
2075.94	Anahola	HI	96703	
5502.26	Ewa Beach	HI	96706	
2602.84	Ewa Beach	HI	96706	

- BusinessType: missing value e.g.

16404	AIEA	HI	96701	531110	Trust
2600	Pearl City	HI	96782	541110	Trust
8500	KAILUA	HI	96734	713990	Trust
130047.5	KIHEI	HI	96753	721191	Trust
56123.05	Kihei	HI	96753	722511	

- JobsReported: missing value

19456	3497.29	Honolulu	HI	96822	812990	Self-Employed Indivi	Unanswered	Unanswered	Unanswered	500	4/30/2020
19457	3256.25	Honolulu	HI	96826	485310	Sole Proprietorship	Unanswered	Unanswered	Unanswered	500	5/1/2020
19458	4367	Aiea	HI	96701	611620	Limited Liability Cor	Unanswered	Unanswered	Unanswered	500	5/7/2020
19459	88100	KAHULUI	HI	96732	423720	Corporation	Unanswered	Unanswered	Unanswered		4/30/2020
19460	4100	WAIPAHU	HI	96797	531390	Corporation	Unanswered	Unanswered	Unanswered		5/14/2020
19461	69300	KAHULUI	HI	96732	561622	Corporation	Unanswered	Unanswered	Unanswered		4/12/2020

notes: we think the business meaning is unclear for value 0 and [blank]. This might need to be explained

- Lender: there are some records having similar names and we suspect it is a data issue. E.g., 1st Financial Bank USA/ First Financial Bank; First Bank / FirstBank

5. Devise an initial plan

- S1: description of dataset D and matching use caseU1;

Dataset D: PPP approved loans from 4/2020 - 8/8/2020 in Hawaii state

Use case U1: we want to know the distribution of average loan amount by cities. And we have checked that the "LoanAmount" column values are all within the range of 50 to 150000. To get the distribution of average loan amount by cities, we need the clean data of the city column.

- S2: profiling of D to identify the quality problems P that need to be addressed to support U1;

Identified data issue in the city column:

a. Some invalid data rows (number, underscore):

LoanAmount	City	State	Zip
44000	5	HI	96814
24300	_	HI	96749

b. Incorrect information (address):

17309.57	1137 11th Avenue	HI	96816
107397.5	184 Puueo Street	HI	96720

c. Upper-case or lower-case city name.

26110	HILO	HI	96720
25948	HILO	HI	96721
25900	HILO	HI	96720
25700	HILO	HI	96720
25680	HILO	HI	96720
25650	HILO	HI	96720
25505	HILO	HI	96720
25412.5	Hilo	HI	96720

Note: from the above data, we also noticed a city can have multiple zip codes.

d. Apostrophe in city names:

10283	Naalehu	HI	96772
5192.5	Naalehu	HI	96772
4167	NAALEHU	HI	96772
1851.87	NAALEHU	HI	96772
1542	NAALEHU	HI	96772
5300	Na'alehu	HI	96772

e. Spaces in city names:

6200	OCEAN VIEW	HI	96737
4832	Ocean View	HI	96737
4100	Oceanview	HI	96737

f. Inconsistent city name:

1338.75	Pearl City	HI	96782
1200	PEARL CITY	HI	96782
1068.28	Pearl City	HI	96782
1031.06	Pearl city	HI	96782
1000	Pearl City	HI	96782
133927	PEARL CITY City	HI	96782
3300	PEARL CITYPearl City	HI	96782
2800	PEARL CITYRL CITY	HI	96782

and

1000	Waipahu	HI	96797
700	WAIPAHU	HI	96797
20900	WAIPAHUAIPAHU	HI	96797
68800	WAIPAHUhu	HI	96797
26000	WAIPAHUhu	HI	96797
7500	WAIPAHUhu	HI	96797
64700	WAIPAHUipahu	HI	96797
65025	WAIPAHUu	HI	96797

We plan to use OpenRefine to address those issues and we can utilize “Zip” column to back check the city value.

- S3: performing the data cleaning process using one or more tools to address the problems P (here you should describe which tools you are planning to use, e.g., OpenRefine;Python; etc.)

Using OpenRefine to clean the data. It is a powerful tool to clean messy data.
By utilizing OpenRefine, we could take advantage of text faceting, titlecase, clustering, trim leading and trailing whitespace functions.

- S4: checking that your new dataset D’ is an improved version of D, e.g., by documenting that certain problems P are now absent and that U1 is now supported;

Plan: for each of those issues we find in step2, we will write a test case to ensure those issues are resolved after cleaning. For the test case steps, we have the following:

Test step1, ensure we can find those issues existed in the raw data.

Test step2, ensure we do not have those issues in the processed data.

- S5: documenting the types and amounts of changes that have been executed on D to obtain D’. You should also include a tentative assignment of tasks to team members (who does what)!

Part 1: Fix structural errors

- ❖ Case consistent for City column (All records will be affected)
- ❖ Fixing the multiple white space for BusinessType column (7140 records will be affected)
- ❖ Same entity with name variance for Lender (4 records will be affected)

Part 2 ： Data validity check

- ❖ Fixing incorrect value for city column (4 records will be affected)

Part 3 ： Perform further cleaning in other fields for potential future use cases

- ❖ Fixing missing values for NaicsCode column (105 records will be affected)
- ❖ Changing blank value to 0 for JobsReported column (2447 records will be affected)
- ❖ Fixing missing value for BusinessType column (1 record will be affected)

Notes: cleaning steps recipe file will be provided for reference purposes.

Team Assignments:

- ◊ Collaboratively we brainstorm on methods and steps of data cleaning.
- ◊ Amy Zhao: S1 & S2
- ◊ Dajun Lin: S3 & S4
- ◊ Xin Peng: S4 & S5