DATA WAREHOUSE PROJECT

Author:

- Magno Alessandro: 4478234

Operational data source:

- “Raise Against Hunger” dataset

Introduction

I chose to answer the following business questions, **Who are their recurring donors? And Do the giving amounts increases or decreases vary based on donor level? What the levels ($ amount cutoffs)? Where do the donations based on recurring donors come from?**

1. Operational data sources inspection and profiling

Looking the data, I tried to understand who are their recurring donors, since there is no definition mentioned for them, I have addressed the analysis of different types of recurring donors based on amount and number of meals donated during the years. I assumed that if an account has donated for two or more times during the years, it is considered as a recurring donor account.

Examples:

1. Type of donor based on amount donated:

- id 22345 donated a total amount of 2500$ in 2010 and a total amount of 3500$ in 2013.
- This person donated money two times in 2010 and in 2013 so he is considered a recurring donor.

2. Type of donor based on number of meals donated:

- id 33456 donated 300 meals in 2006 and 100 meals in 2010.
- This person donated meals two times in 2010 and in 2013 so he is considered a recurring donor.

Since the donor levels are not specified, I have considered the following donor levels for my analysis by looking the values of amount and number of meals donated during the years:

LEVEL 1: 1 TO 1000

LEVEL 2: 1000 TO 10000

LEVEL 3: 10000 TO 50000

LEVEL 4: 50000 TO 100000

LEVEL 5: 100000 TO 1000000

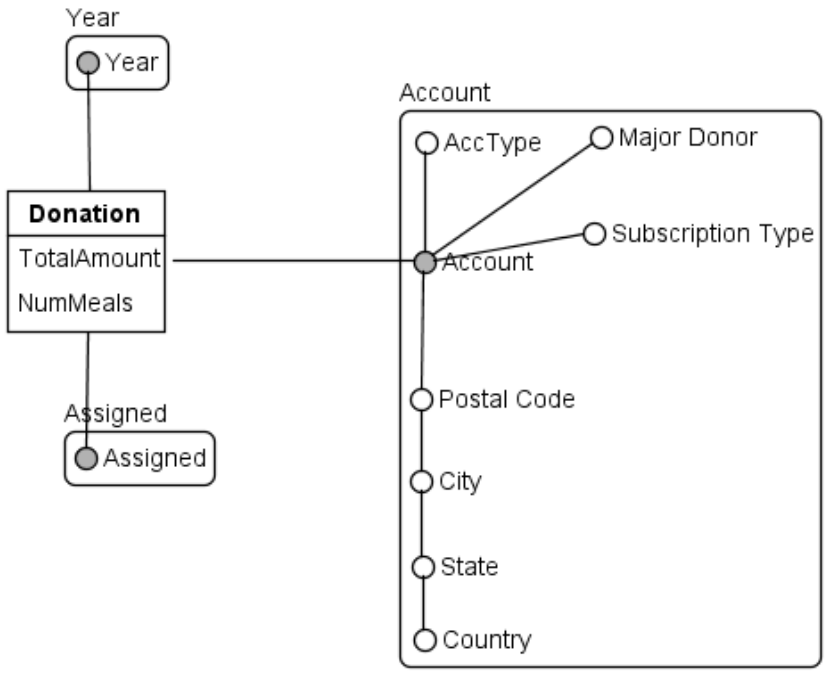
LEVEL 6: 1000000 AND ABOVE

Before starting the data cleaning, I asked myself:

- what are the columns that are required to answer the questions?
- are the columns understandable using data dictionary provided?

I tried to understand the meaning of the various columns and I extracted some of them using a python script then I imported them in the Tableau Prep Builder for starting the cleaning.

2. Data warehouse conceptual design



COMMENTS AND EXPLANATIONS

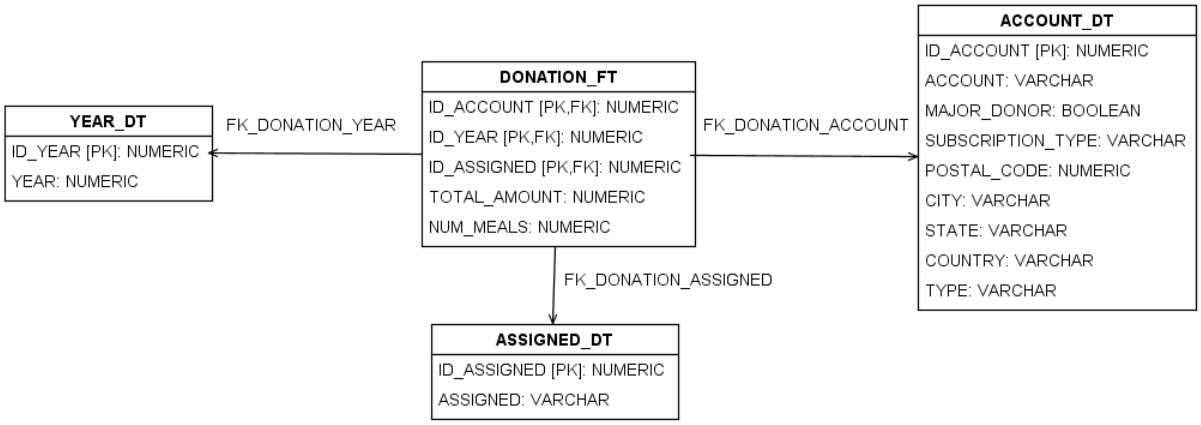
The questions that I want to answer are all about donations, so it seems reasonable to consider a donation as a fact. I want to measure the total amount and the number of meals donated by donors, year and assigned (the donors that referring other donors). Talking about dynamicity in dimensions, I have one hierarchy and the attributes values that populating it are static because they refer to geographic fields.

Why the conceptual schema is not so articulated?

The reasons are two:

1. I struggled with the cleaning of data, so I have to reduce the business questions to focus on less columns and this helped me to use Tableau Prep Builder without crashes and without losing my work.
2. Linked to I have just said in the first point, focusing on less columns (less data), I was not able to do a lot of dimensions and hierarchies. So I picked those that helped me to answers the business questions.

3. Data warehouse ROLAP design



Volume estimation

Data volumes:

Number of facts: $2*10^6$

account: $16*10^4$

subscription_type: 25

postal_code: $4*10^4$

state: 60

city: $2*10^4$

country: 1

type: 9

year: 11

assigned: 60

I assumption that the workload consists of aggregated queries on the following patterns:

- Q1 = {state, year}
- Q2 = {city, year}

Which one could be selected as a candidate key?

Q1 = 60*11 = 660

Q2 = $2*10^4$ = 20000

It is worthwhile materializing Q1 since it is small compared to Q2.

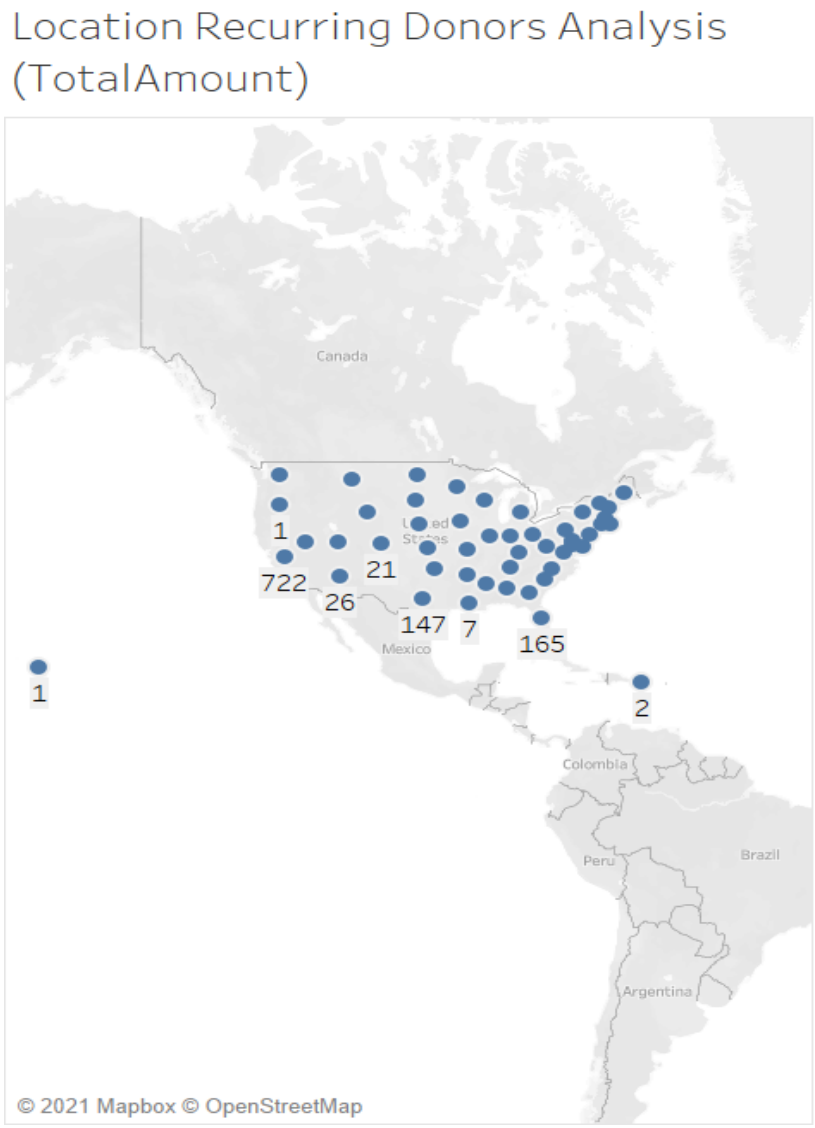
ETL

As I said above, I had problems with Tableau Prep Builder for cleaning my data, so I decided to extract the columns that I needed for answering the business questions. I have done that with a script in python using library pandas. After that I proceeded with the cleaning using Tableau Prep Builder, looking the data I saw that the geographic fields were very noisy. I only considered the records having US as Country because it was more than 50% of the dataset which is a considerable amount of data for analysis. I fixed the postal code, city and state using the data set from the following link: http://federalgovernmentzipcodes.us/

Then I imported the file in PostgreSQL where I have done some analysis for answering the questions and defined my fact and dimension tables.

6. Tableau

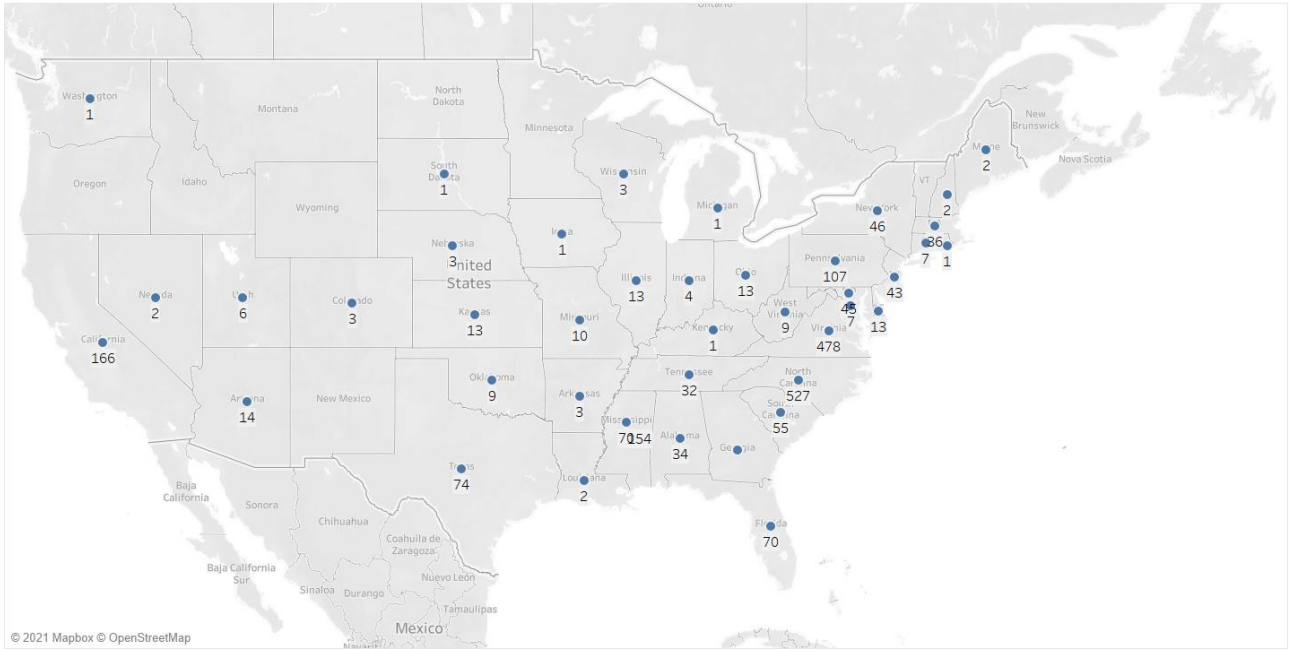
Data Analysis and Visualization



- This recurring donor analysis is conducted based on how many times they have donated (total amount) from year 2010 to 2016.
- If an account has donated for two or more times, it is considered as recurring donor account.

As it can be seen from the image above, most of the donors are in East Coast.

Location Recurring Donors (Meals Donated)

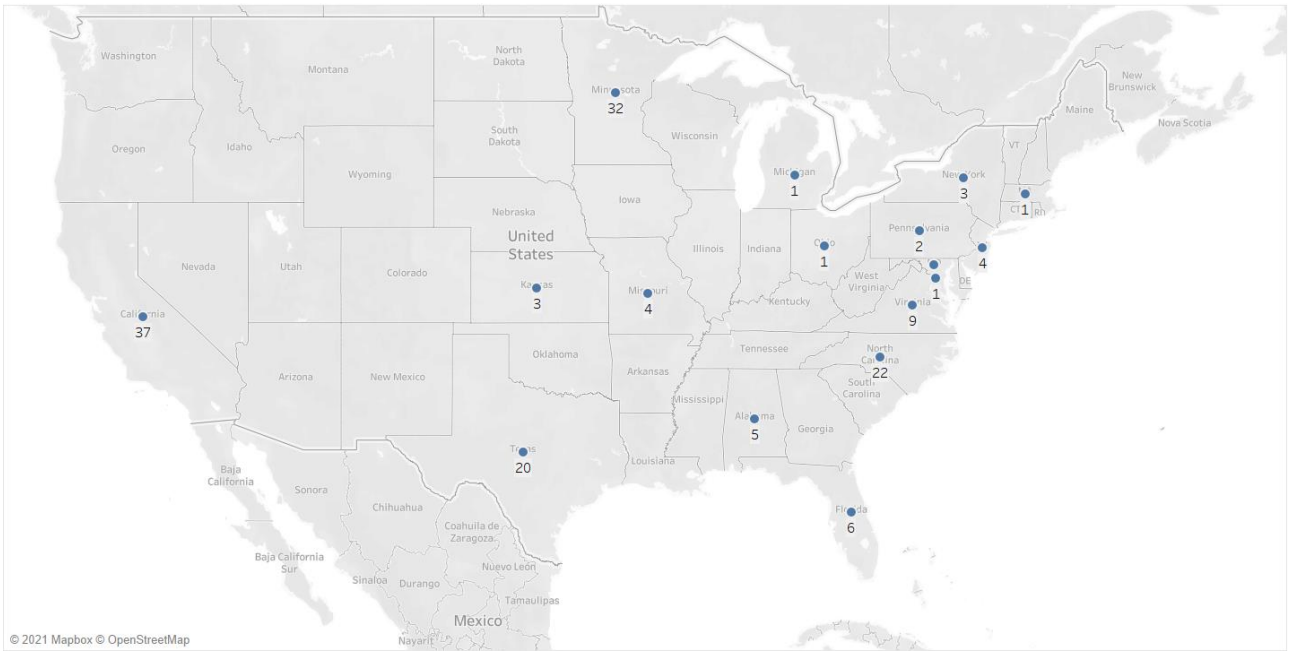


- This recurring donor analysis is conducted based on how many times they have donated meals from year 2010 to 2016.
- If an account has donated for two or more times, it is considered as recurring donor account.

As it can be seen from the image above, most of the donors are in East Coast.

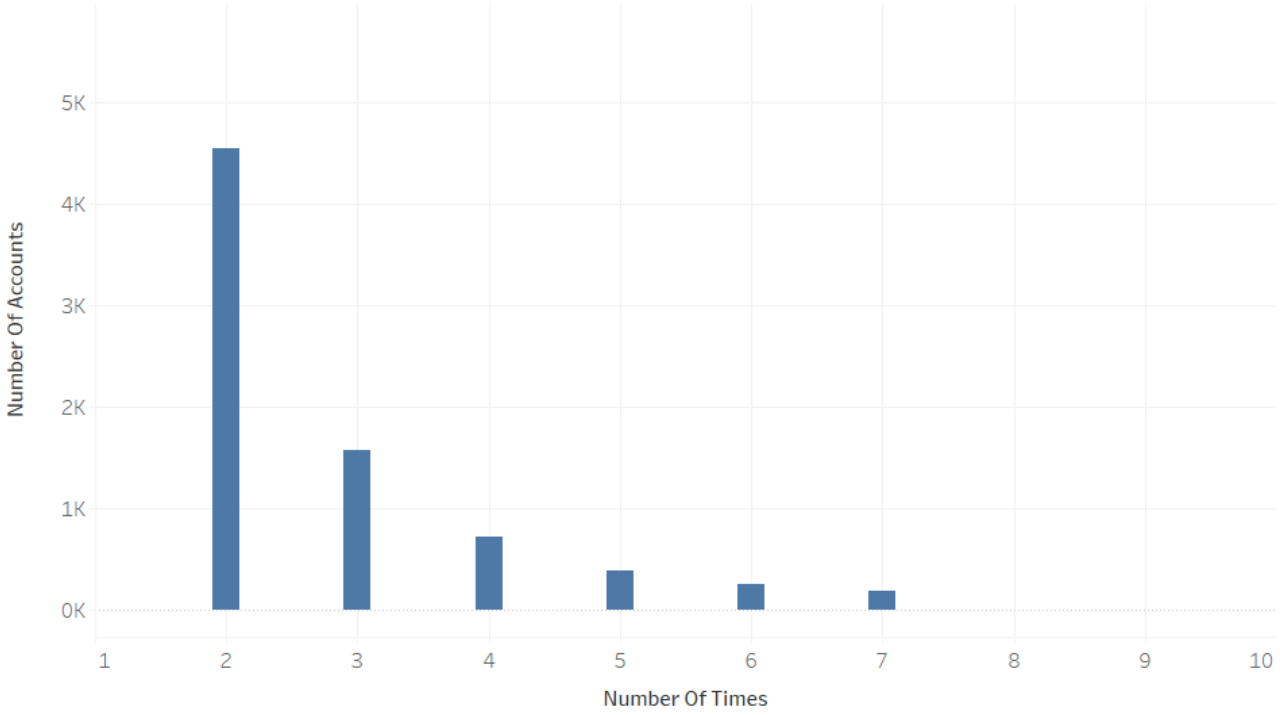
On the west coast just California as a good amount of donors.

Location Major Donor Analysis



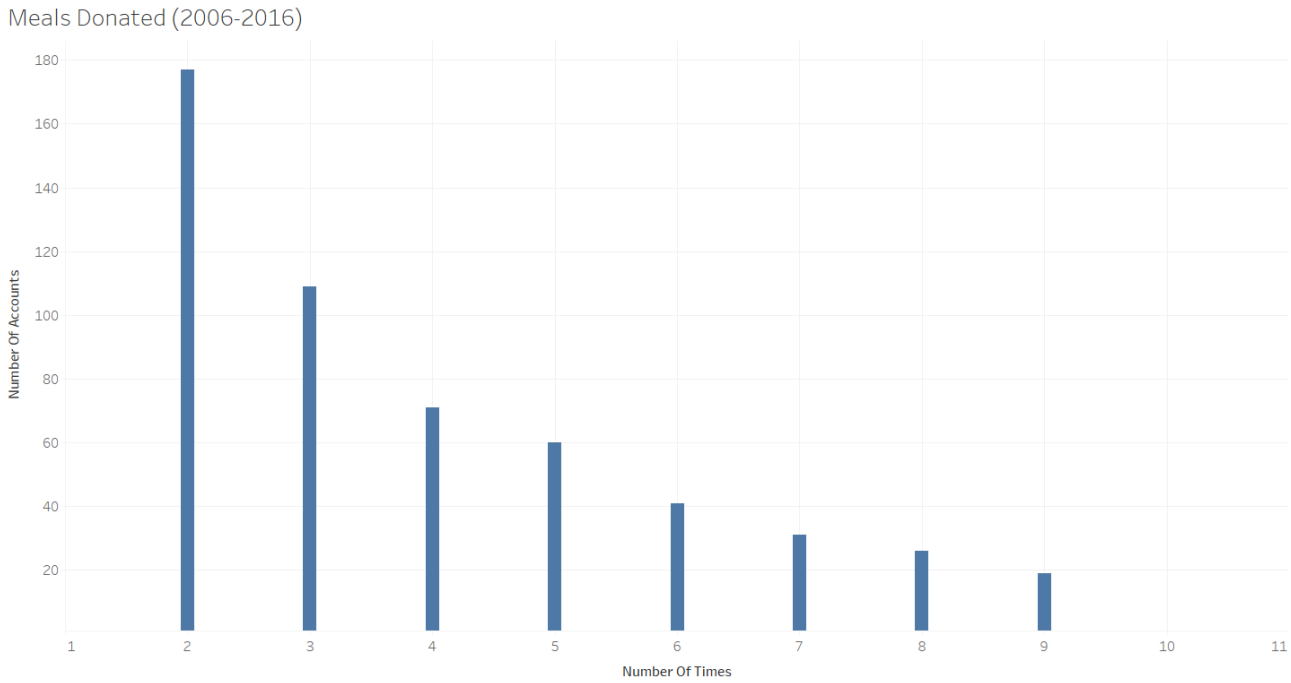
- Most of the major donors are located on the East Coast of the USA

Amount Donated (2010-2016)

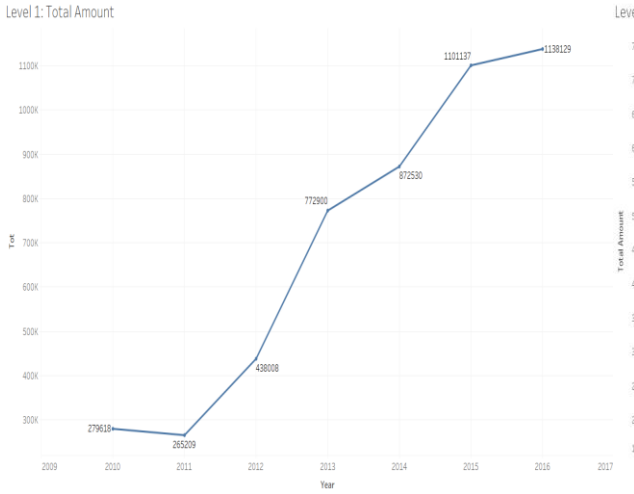


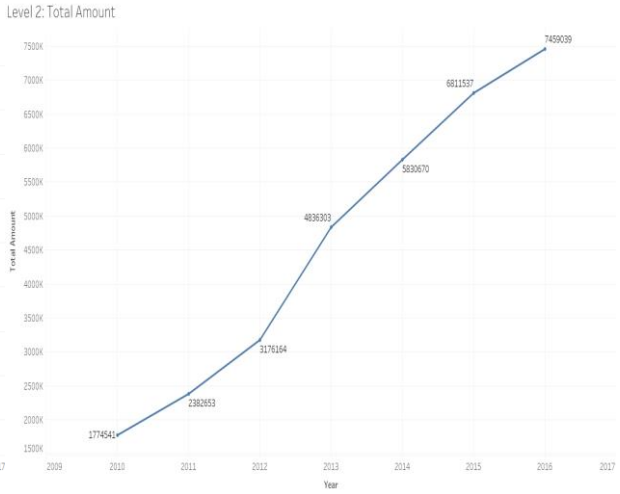
- Analysis of total amount donated based on number of accounts vs the frequency of donation.
- This analysis is done to analyze who are the loyal donors and how many times they have donated.

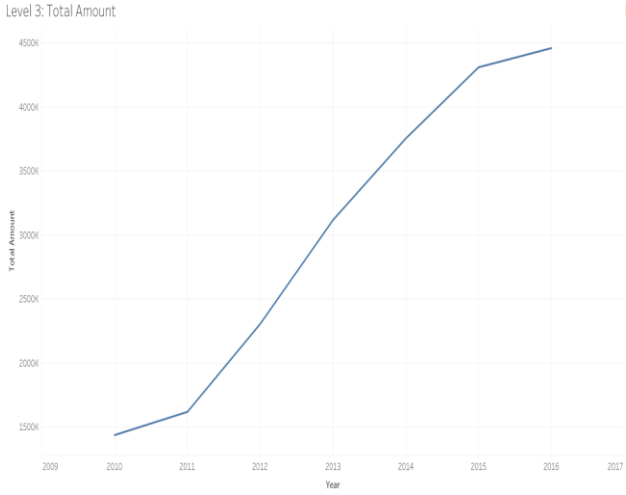
- There are a few donors who have been there with the organization for almost all the years.

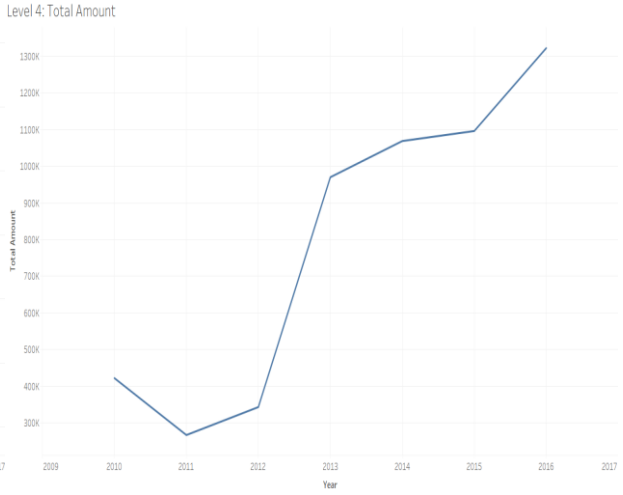


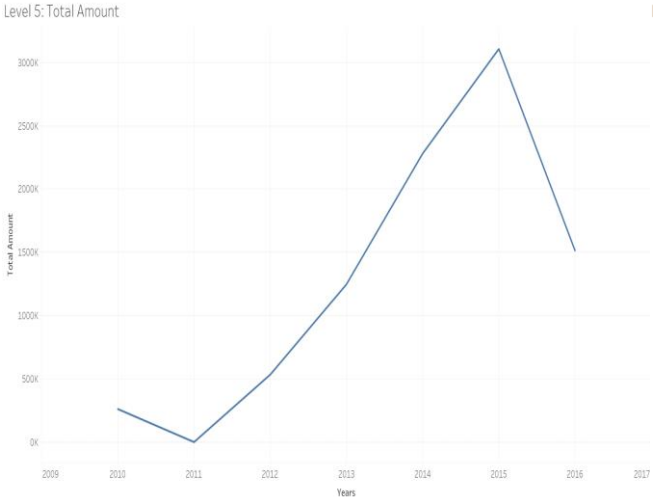
- Analysis of meals donated based on number of accounts vs the frequency of donation.
- This analysis is done to analyze who are the loyal donors and how many times they have donated.
- There are a few donors who have been there with the organization for almost all the years.

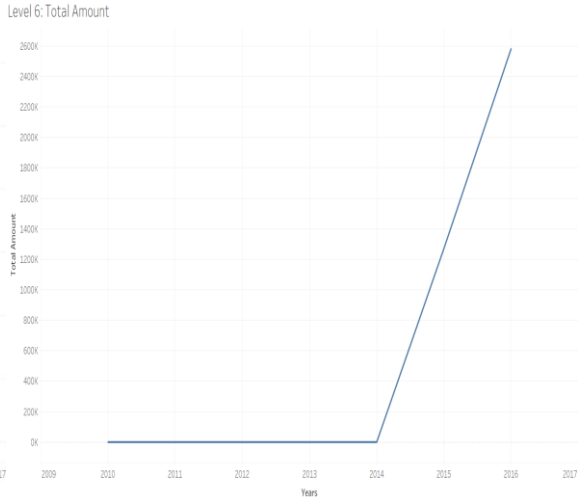












The plots above show how the total amount is changing (increasing or decreasing) based on each Donor Level.

- The overall trend of donation is increasing for most of the donor levels.
- There were no donations in Level 6 till 2014, but after that the donations have increased from year 2015 to 2016.

- The level 5 donations have taken a dip from 2015 to 2016. Since the donation amount is very high in this level, if one or two donors decide not to donate, total donation will decrease by significant amount.

Tools used:

- Tableau Prep Builder
- Tableau Desktop
- Python Library Pandas
- PostgreSQL
- BI Modeler

The tools that gave me more problems is Tableau Prep Builder although is very useful for cleaning the data but too much data made it crashed a lot of times. About pandas library in python is very efficient to read csv and to extract columns. I have to valuate in a positive way PostgreSQL, very good for cleaning and analyze data, also Tableau Desktop for visualization and analyze data worked well. BI Modeler is the software that I used to build the conceptual schema and rolap schema because I had some problems to download Indyco. That is a free software made by an Italian professor Stefano Cazzella, the problem that I had regarding the lack of documentation.

Effort of the project

The time dedicated is about 90 or more hours. The reason is the problems with tableau prep builder that make me lost a lot of work so I have done it again. Another problem was passing the data from tableau prep builder to PostgreSQL. Another one is the conceptual schema that I have done the first time, it had more dimensions and hierarchies but it didn’t make sense because in the data were not matches, so I had to modify it and adjust the code.