MSDA 607 Final Project

Ben Arancibia

December 9, 2014

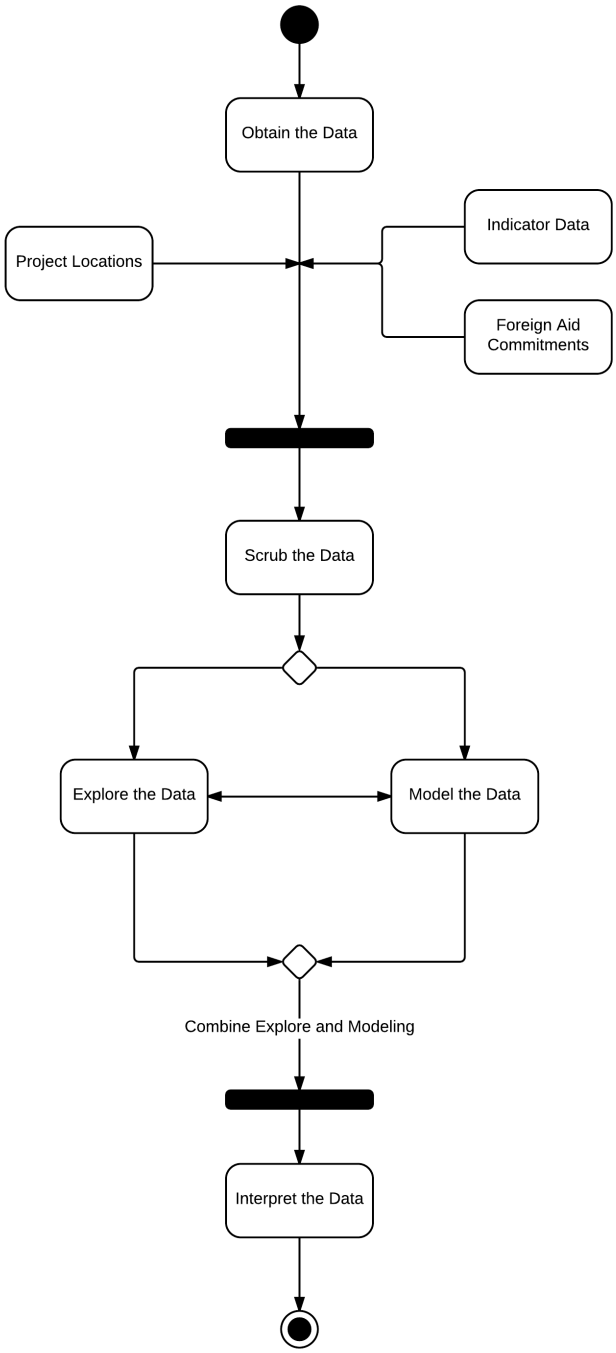
Background

This is report is the MSDA 607 - Data Acquistion and Management Final Project. As detailed in the project proposal three different datasets will be used to look at financial foreign aid in Honduras. The three datasets are Honduras foreign aid commitments, Honduras financial foreign aid project's locations, and an indicator level dataset with various types of data. This project will be broken into five sections as follows: 1) Introduction and Workflow, 2) Obtaining the Data, 3) Data Transformation, 4) Statistical Analysis and Visualization, 5) Conclusion. Most of the data was collected via the Aid Management Platform (AMP) which my company builds for developing countries to record their foreign aid. This project is important and has real world applications because the results will be used to continue my company's transition from data collectors towards data analysis.

Introduction and Workflow

This project is using three different datasets created by different organizations. The Honduras Foreign Aid Commitments was entered into the AMP by the Government of Honduras (GoH), the AMP data is stored in a PostGres Database. The locations data for each project was obtained from the Donors that fund projects in Honduras. This location data is not shared with the GoH and is one of the many pitfalls of the international development sector. The indicator level dataset which has data from the municipalities of Honduras including population density and number of Hospitals per municipality was created by a GoH Ministry called El Sistema Nacional de Información Territorial which has been since dissolved because of elections last year.

The data science workflow of this project will follow a typical OSEMN workflow, which can be seen below.



This workflow will be used during this analysis.

Obtain the Data

The data was obtained from three different sources. The Honduras Foreign Aid Projects are located in a Postgres Database. Unfortunately do to security reasons this report cannot have a direct query to the database, but the query used can be shown and how a user would connect to PostGreSQL Database from R if desired (for this project the command line was used).

```
library(sqldf)

library(RPostgreSQL)

drv <- dbDriver("PostgreSQL")
db <- dbConnect(drv, dbname="Honduras",
                 user="username", password="password")


sqldf("SELECT *
      FROM
      view_todos_proyectos") # view for all projects in the database

sqldf("COPY (SELECT *
      FROM view_todos_proyectos)
      TO '/users/bcarancibia/CUNY_IS_607/Final_Project'
      USING DELIMITERS ',' WITH  CSV") # copy for all projects in the database
```

I exported the data using a view because there are about ~150 tables in the database with different data, it is just easier to select the view and export. To get the view a user would need to query and join about 45 tables to get all the different variables shown in an entire db export, this is something currently being streamlined during a reengineering process. This data is completely open and can be downloaded form the Honduras AMP website www.pgc.sre.gob.hn.

The other two datasets were downloaded in CSV format. The locations data is downloaded from the GoH AMP Website (www.pgc.sre.gob.hn). The SINIT dataset which is the indicator dataset was downloaded from Development Gateway's internal repository of data Wiki.

The three datasets are then imported into R.

```
honduras_indicator = read.csv("/users/bcarancibia/CUNY_IS_607/Final_Project/indicator_honduras.csv",
                     header = TRUE)

honduras_projects = read.csv("/users/bcarancibia/CUNY_IS_607/Final_Project/commitments_honduras.csv",
                     header = TRUE)

honduras_locations = read.csv("/users/bcarancibia/CUNY_IS_607/Final_Project/locations_honduras.csv",
                     header = TRUE)
```

Always good to take a look at the data after import to make sure everything went well.

```
head(honduras_indicator) #indicator

head(honduras_projects) #project commitments

head(honduras_locations) #locations of projects
```

After performing this there are about 15 pages of data to be shown which is a lot for three datasets and this size is not the focus of this report. It also means there is a lot of data that can be scrubbed and trimmed down.

Data Transformation

Now that the data has been imported a couple data transformations need to be done. The projects and locations need to be joined by the ID known as AMPID. After joining these two files together a new file will be created with just the columns needed to do analysis. This new file will then be joined with indicator data by municipality.

The first step is to a subset of the Honduras Projects dataset. If you were to look at Honduras Projects view, you would see that there are 185 variables. 185 variables makes it very complicated to look at the data and joining it with another file of will be difficult and cumbersome. For this report the following fields will be selected:

- AMPID
- Sector.Ejecutor (Executing Sector)
- Nombre.de.Proyecto.Programa (Name of the Project/Program)
- Agencia.financiera (Financing Agency)
- Grupo.Donante (Donor Group)
- Socio.al.desarrollo.Responsable (Responsible Development Group)
- All Commitment Months by Year

Subset below:

```
hn_project_subset <- honduras_projects[c(1,5,6,13,16,26,39:185)]
hn_project_subset[is.na(hn_project_subset)] <- 0 #make NAs = 0
```

Now that there is a subset of data, the Honduras Project Subset will be joined with the Honduras Locations.

```
project_locations <- merge(hn_project_subset, honduras_locations, by="AMPID")
```

Looking at the successful merge there are now some repeated fields, i.e. "TITLE" is the same as Name of Project/Program. For this report, a final subset will be done in order to create the final dataset to be analyzed and visualized.

Final Subset below:

```
hn_projects_locations <- project_locations[c(1:153, 155, 156, 157)]
```

Statistical Analysis and Visualization

Now that the data is in a final version, it makes sense to do a summary analysis on some of the data fields. One data field to do a summary of is the Total data field, which is the total amount of commitments (money) per project over the entire project. Many times projects last multiple years.

```
summary(hn_projects_locations$Total) #summary stats of projects
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##         0   2545000  10070000  19400000  32800000 300000000
```

```
sd(hn_projects_locations$Total) #standard deviation of projects
```

```
## [1] 20776332
```

The standard deviation for total projects is 20,776,332 USD. That is a large standard deviation. Create a new field that is aggregate of each total by sector. Using ggplot2 plot the totals of each sector of foreign aid projects.

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
by_sector <- group_by(hn_projects_locations, TYPE)
summary_by_sector <- summarise(by_sector, count_each_sector = n(),
                               total_by_sector = sum(Total))

order_summary_sector <- summary_by_sector[order(-summary_by_sector$total_by_sector),]

order_summary_sector
```

```
## Source: local data frame [19 x 3]
##
##                                                TYPE count_each_sector
## 1                       MULTISECTOR / CORTE CRUZADO              1097
## 2                           TRANSPORTE Y ALMACENAJE               293
## 3                                       AGRICULTURA               313
## 4                  GENERACION Y FUENTE DE LA ENERGIA               315
## 5           OTRA INFRAESTRUCTURA Y SERVICIOS SOCIALES               678
## 6                                        EDUCACION               205
## 7                         GOBIERNO Y SOCIEDAD CIVIL               428
## 8                               AGUA Y SANEAMIENTO               191
## 9                                           SALUD                88
## 10 AYUDA DE LA MATERIA Y AYUDA GENERAL DEL PROGRAMA                36
## 11                                         #NAME?                29
## 12                                      INDUSTRIA                63
## 13                                AYUDA HUMANITARIA                22
## 14                         ACCION REFERENTE A DEUDA                 1
## 15                                          PESCA               103
## 16                                    COMUNICACION                 1
## 17                                    SILVICULTURA                15
## 18                                  332 - TURISMO                 5
## 19       PROGRAMAS Y POLITICAS DE POBLACION & SALUD                 1
## Variables not shown: total_by_sector (dbl)
```

As seen in ordered by aggregateed sector there are 19 different sectors with the largest amount of foreign aid going to Multisector and the smallest amount of aid going to Programas y Politicas de Poblacion y Salud

(Programs and Politics of Health and Population). The sectors best defined and receive the most amount of aid are Transporte y Almacenaje and Agriculture. This makes sense because of Hurricane Mitch in the late 1990s wiped out about 80% of Honduras' infrasture and the country is still trying to recover from that hurricane. Also an interesting thing to look at in the data is that there is a sector without a named defined. There still obvious data clean up and better data management that needs to be done in-country.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
f1 <- ggplot(order_summary_sector, aes(x=order_summary_sector$TYPE, y=order_summary_sector$total_by_sec
  geom_histogram()
```

Conclusion