Cleaning and Storing Supply Chain Delivery Data

DSCS6020 Collect/Store/Retrieve Data

Abhiram Paruchuri

100933135

Introduction

Project Proposal:

The proposed project idea is to clean a supply chain delivery data set, obtained from The United States President's Emergency Plan for Aids Relief (pepfar). The organization's main goal is to provide better health care and testing facilities to the places that are lacking them. The data consists of all the deliveries made by the pepfar delivery network over the past few years. Cleaning and preprocessing the data can help in using the data to observe trends in the medical needs and help in effectively predicting the inventory. As supply chain is a constantly evolving process, effectively forecasting the changes and trends can help in developing a better plan to handle the needs. In this case, it is the HIV medications and test kits. In addition, storing the dataset into a database can make visualizing the data easier.

In addition, I have also retrieved the HIV/AIDS prevalence rates, deaths caused by AIDS, number of people infected from the data on the size of HIV epidemic from WHO global health observatory data repository. The idea is to collect all the available data on the epidemic and combine it into a single repository. This will make analysis easier and improve the scope of the supply chain network by providing answers to the questions like, which is the most affected country?, how many are in need of medical aid? Etc.

Brief introduction to prepfar:

The U.S. President's Emergency Plan for AIDS Relief (PEPFAR) is the U.S. Government initiative to help save the lives of those suffering from HIV/AIDS around the world. This historic commitment is the largest by any nation to combat a single disease internationally, and PEPFAR investments also help alleviate suffering from other diseases across the global health spectrum. PEPFAR is driven by a shared responsibility among donor and partner nations and others to make smart investments to save lives.

PEPFAR is the cornerstone and largest component of the U.S. President's Global Health Initiative. With a special focus on improving the health of women, newborns and children, the Global Health Initiative's goal is to save the greatest number of lives by increasing and building upon what works and, then, supporting countries as they work to improve the health of their own people.

Brief introduction to SCMS:

The SCMS aim is to provide cost-effective, reliable, secure and sustainable supply chains for developing countries, which can be a huge thing for people suffering with HIV. For more than six years, the Supply Chain Management System (SCMS) has been saving lives through stronger supply chains. In collaboration with in-country and international partners, SCMS:

- Provides quality, best-value health care products to those who need them.
- Deploys innovative solutions to help programs enhance their supply chain capacity.
- Ensures accurate supply chain information is collected, shared and used.

Cleaning the Data

Dataset Overview:

There are two data sets that are being derived. They are:

1. The SCMS scheduled delivery data set
2. Dataset created from WHO country statistics.

The SCMS scheduled delivery dataset:

First, we deal with the SCMS scheduled delivery data set, which comprises of procurement transaction data from the Supply Chain Management System (SCMS), administered by the United States Agency for International Development (USAID), i.e. data about all the deliveries made, which are related to medical needs and facilities like, drugs, tests and suspension kits.

The dataset captures deliveries of antiretroviral (ARV)drugs, rapid diagnostic tests for HIV and malaria , and anti-malaria medicines, including prices and delivery destinations. This data is particularly valuable for understanding ranges and trends in pricing as well as

First we will review the glossary of the data set and talk about the problems we have with it.

ID	FieldName	FieldDescription
1	ID	Primary key identifier of the line of data in our analytical tool
2	Project Code	Project code
3	PQ #	Price quote (PQ) number
4	PO #	Order number: Purchase order (PO) for Direct Drop deliveries, or Sales Order (SO) for from Regional Delivery Center (RDC) deliveries
5	ASN/DN #	Shipment number: Advanced Shipment Note (ASN) for Direct Drop deliveries, or Delivery Note (DN) for from RDC deliveries
6	Country	Destination country
7	Managed By	SCMS managing office: either the Program Management Office (PMO) in the U.S. or the relevant SCMS field office
8	Fulfill Via	Method through which the shipment was fulfilled: via Direct Drop from vendor or from stock available in the RDCs
9	Vendor INCO Term	The vendor INCO term (also known as International Commercial Terms) for Direct Drop deliveries
10	Shipment Mode	Method by which commodities are shipped
11	PQ First Sent to Client Date	Date the PQ is first sent to the client
12	PO Sent to Vendor Date	Date the PO is first sent to the vendor

13	Scheduled Delivery Date	Current anticipated delivery date
14	Delivered to Client Date	Date of delivery to client
15	Delivery Recorded Date	Date on which delivery to client was recorded in SCMS information systems
16	Product Group	Product group for item, i.e. ARV, HRDT
17	Sub Classification	Identifies relevant product sub classifications, such as whether ARVs are pediatric or adult, whether a malaria product is an artemisinin-based combination therapy (ACT), etc.
18	Vendor	Vendor name
19	Item Description	Product name and formulation from Partnership for Supply Chain Management (PFSCM) Item Master
20	Molecule/Test Type	Active drug(s) or test kit type
21	Brand	Generic or branded name for the item
22	Dosage	Item dosage and unit
23	Dosage Form	Dosage form for the item (tablet, oral solution, injection, etc.).
24	Unit of Measure (Per Pack)	Pack quantity (pills or test kits) used to compute unit price
25	Line Item Quantity	Total quantity (packs) of commodity per line item
26	Line Item Value	Total value of commodity per line item
27	Pack Price	Cost per pack (i.e. month's supply of ARVs, pack of 60 test kits)
28	Unit Price	Cost per pill (for drugs) or per test (for test kits)
29	Manufacturing Site	Identifies manufacturing site for the line item for direct drop and from RDC deliveries
30	First Line Designation	Designates if the line in question shows the aggregated freight costs and weight associated with all items on the ASN/DN
31	Weight (Kilograms)	Weight for all lines on an ASN/DN
32	Freight Cost (USD)	Freight charges associated with all lines on the respective ASN/DN

33	Line Item Insurance (USD)	Line item cost of insurance, created by applying an annual flat rate (%) to commodity cost

Problems faced with the dataset:

As this is the data that is made available for public, there are some key limitations to the data. Multiple deliveries are consolidated into single delivery, so freight cost is inaccurate relative to the line item. The dates are not properly documented. As supply chain data vary for different transactions and the data changes overtime. So NA values are a problem. Some dates are not applicable as the goods for those order are fulfilled out of stock available at RDC (Regional Development Center).The Line Item Value/Pack Price/Unit Price values have high variability as they can be sometimes low due to donations.

Cleaning the data:

First we clean the columns with the date values. U sing the lubridate package, we convert all the values in the date columns to date classes. The columns (PO.Sent.to.Venfor.Date, Sheduled.Delivery.Date, Delivered.to .Client.Date) are converted to numeric values, so as to find the delivery time and the delay time by subtracting the PO sent date and scheduled date from the delivered date. If the delayed time value is a negative vale, then the product arrived earlier than expected.

In the program, it is done by constructing a deriveDates() function

Screen shots:

Result:

PO.Sent.to.Vendor.Date	Scheduled.Delivery.Date	Delivered.to.Client.Date	Delivery.Recorded.Date
Date Not Captured	2-Jun-06	2-Jun-06	2-Jun-06
Date Not Captured	14-Nov-06	14-Nov-06	14-Nov-06
Date Not Captured	27-Aug-06	27-Aug-06	27-Aug-06
Date Not Captured	1-Sep-06	1-Sep-06	1-Sep-06
Date Not Captured	11-Aug-06	11-Aug-06	11-Aug-06
Date Not Captured	28-Sep-06	28-Sep-06	28-Sep-06
Date Not Captured	8-Jan-07	8-Jan-07	8-Jan-07
Date Not Captured	24-Nov-06	24-Nov-06	24-Nov-06
Date Not Captured	7-Dec-06	7-Dec-06	7-Dec-06
11/13/2006	30-Jan-07	30-Jan-07	30-Jan-07
12/1/2006	16-Feb-07	16-Feb-07	16-Feb-07
Date Not Captured	8-Jan-07	8-Jan-07	8-Jan-07
Date Not Captured	10-Jan-07	10-Jan-07	10-Jan-07
12/22/2006	27-Feb-07	27-Feb-07	27-Feb-07
Date Not Captured	18-Jan-07	18-Jan-07	18-Jan-07
1/10/2007	19-Mar-07	19-Mar-07	19-Mar-07
Date Not Captured	7-May-07	7-May-07	7-May-07
Date Not Captured	29-Mar-07	29-Mar-07	29-Mar-07
4/12/2007	6-Jun-07	6-Jun-07	6-Jun-07
5/13/2007	19-Jun-07	19-Jun-07	19-Jun-07
5/17/2007	19-Jun-07	19-Jun-07	19-Jun-07
7/13/2007	2-Oct-07	2-Oct-07	2-Oct-07
7/4/2007	15-Oct-07	15-Oct-07	15-Oct-07

The dates before cleaning

PO_Sent_to_Vendor_Date	Scheduled_Delivery_Date	Delivered_t_Client_Date	Year	Delivery_Time	Delay_Time
NA	2006-06-02	2006-06-02	2006	0	0
NA	2006-11-14	2006-11-14	2006	0	0
NA	2006-08-27	2006-08-27	2006	0	0
NA	2006-09-01	2006-09-01	2006	0	0
NA	2006-08-11	2006-08-11	2006	0	0
NA	2006-09-28	2006-09-28	2006	0	0
NA	2007-01-08	2007-01-08	2007	0	0
NA	2006-11-24	2006-11-24	2006	0	0
NA	2006-12-07	2006-12-07	2006	0	0
2006-11-13	2007-01-30	2007-01-30	2007	78	0
2006-12-01	2007-02-16	2007-02-16	2007	77	0
NA	2007-01-08	2007-01-08	2007	0	0
NA	2007-01-10	2007-01-10	2007	0	0
2006-12-22	2007-02-27	2007-02-27	2007	67	0
NA	2007-01-18	2007-01-18	2007	0	0
2007-01-10	2007-03-19	2007-03-19	2007	68	0
NA	2007-05-07	2007-05-07	2007	0	0
NA	2007-03-29	2007-03-29	2007	0	0
2007-04-12	2007-06-06	2007-06-06	2007	55	0
2007-05-13	2007-06-19	2007-06-19	2007	37	0
2007-05-17	2007-06-19	2007-06-19	2007	33	0
2007-07-13	2007-10-02	2007-10-02	2007	81	0
2007-07-04	2007-10-15	2007-10-15	2007	103	0
2007-07-04	2007-08-27	2007-08-27	2007	54	0
2007-07-26	2007-08-13	2007-08-21	2007	26	8
2007-07-26	2007-08-25	2007-08-25	2007	30	0
NA	2007-10-16	2007-10-16	2007	0	0
NA	2007-11-22	2007-11-22	2007	0	0
NA	2007-11-22	2007-11-22	2007	0	0
2007-10-03	2007-11-20	2007-11-20	2007	48	0
2007-08-28	2007-10-03	2007-10-03	2007	36	0
2007-11-12	2008-01-29	2008-01-29	2008	78	0
2007-11-19	2008-01-21	2008-01-21	2008	63	0
2007-11-21	2008-01-21	2008-01-21	2008	61	0
2007-12-10	2008-01-31	2008-01-31	2008	52	0
NA	2008-02-05	2008-02-05	2008	0	0
2008-01-04	2008-01-21	2008-01-04	2008	0	-17

The cleaned dataset for dates

Assigning the Product name:

Now we assign a proper name to the product from the 'Molecule Test Type' column and the Brand name column. We parse through both the columns, and choose the brand name when it is present and choose the molecule name when the value is listed as generic as it means that the molecule name is generally the product name.

Result:

Molecule.Test.Type	Brand
HIV, Reveal G3 Rapid HIV-1 Antibody Test	Reveal
Nevirapine	Generic
HIV 1/2, Determine Complete HIV Kit	Determine
Lamivudine	Generic
Stavudine	Generic
Zidovudine	Generic

A .The above figure has values before merging

Product_Name
Reveal
Nevirapine
Determine
Lamivudine
Stavudine
Zidovudine
Stocrin/Sustiva
Nevirapine

B. values after merging

Splitting the drug into its parts:

Several drugs contain two or more types of molecules in them. By constructing the splitElements() function, we split the Molecules and dosages into separate columns. The use of splitting the drugs into their classes is to increase the scope of analysis on the drug type and combination.

Result:

Efavirenz/Emtricitabine/Tenofovir Disoproxil Fumarate	Atripla	600/200/300mg
Quinine (as dihydrochloride)	Generic	600mg/2ml
HIV 1/2, Uni-Gold HIV Kit	Uni-Gold	N/A
Lamivudine/Zidovudine	Generic	150/300mg
Lamivudine	Generic	150mg
Lamivudine/Nevirapine/Stavudine	Generic	150/200/30mg
Lamivudine/Zidovudine	Generic	150/300mg
Lopinavir/Ritonavir	Kaletra	80/20mg/ml
Lopinavir/Ritonavir	Generic	100/25mg

The above table consists of molecule names and drug dosages before cleaning

Lopinavir	Ritonavir	NA
Malaria Paramax-3 Kit	NA	NA
Efavirenz	Emtricitabine	Tenofovir Disoproxil Fumarate
Quinine (as dihydrochloride)	NA	NA
test	NA	NA
Lamivudine	Zidovudine	NA
Lamivudine	NA	NA
Lamivudine	Nevirapine	Stavudine
Lamivudine	Zidovudine	NA
Lopinavir	Ritonavir	NA
Lopinavir	Ritonavir	NA
test	NA	NA

Molecules split into their respective columns

263.40	200.0	50.0
158.05	300.0	0.0
68.63	30.0	0.0
13.72	0.0	0.0
824.89	200.0	50.0
30.07	100.0	0.0
2.12	80.0	20.0
5.52	300.0	200.0
38.75	300.0	0.0
41.16	0.0	0.0
1176.00	0.0	0.0
1002.74	0.0	0.0
757.31	600.0	0.0
215.60	0.0	0.0
1.20	100.0	0.0
76.49	150.0	300.0

Dosages split into their respective columns

Assigning the dosages to their respective units & donation status:

As many data analysis tools are sensitive to units of measure, we split the dosages in the column, which are jumbles into mg/ml, mg and g values to their respective unit columns

We also construct a new column to decide if the line item is a donation or not. The condition set is, if the pack price is less than 2.5 dollars, we designate it as a donation.

Result:

Dosage
N/A
10mg/ml
N/A
150mg
30mg
10mg/ml
200mg
200mg
30mg

Dosage Column before

Dosage_ml_mg	Dosage_mg
NA	NA
10	NA
NA	NA
NA	150
NA	30
10	NA
NA	200
NA	200
NA	30
NA	200/50
NA	200/50
NA	NA
NA	NA
NA	150/300

After

Donation_Designation
N
N
N
N
N
N
N
N
Y
N
N
N

Donation Column

Final note:

Finally the cleaned data is loaded into a data frame that can be loaded into a database. Some of the columns with redundant values and useless columns are left out while populating the new data frame as they are considered less important for data analysis or further inference.

Not considering the columns above, the other columns that are considered are, ID, project code, PO.SO, ASN, Country, Managed, Fulfill via, shipment mode, vendor, unit of measure per pack, Item Quantity, Pack price, first line designation, weight in kilograms, insurance.
The NA values are converted to zeros rather than omitting as they are very widely dispersed.

Second dataset retrieval:

The second data set is constructed using three datasets that are downloaded in program from the WHO Global Health Observatory Data Repository. The cleaning is done in the following steps:

a. The data sets for country wise HIV prevalence, deaths and number of people infected are downloaded and opened in R in the form of a data frame.
b. We observe that all the columns are polluted with '&lt' values, which might have occurred during retrieving.
c. We parse through each values in the three data frames and remove the '&lt' values from the values.
d. Now we notice that the values for the prevalence, deaths and the number of people infected and their confidence interval values are present in the same columns.
e. We construct a function numSplit() to separate these values into two separate columns, while removing the white spaces and the square brackets present in the value.
f. Now we merge all the three datasets in to a single data set using merge() function. First prevalence and deaths data are merged on prevalence country column and then the merged data is merged with number of infected people data set with respect to the same prevalence country column so that we have a common binder.
g. NA values are removed using Na.omit() function as they are not dispersed , but are present in the form of rows.
h. All the data cleaning is done by the dataClean() function the code. The rename() function renames the data after cbind.

Screen shots:

The screen shots are provided only for the prevalence data set. It is the same for the other data sets also.

Country	X.2013	X.2009	X.2005	X.2001
Afghanistan	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]
Albania	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]	<0.1 [<0.1-0.1]
Algeria	0.1 [0.1-0.2]	0.1 [<0.1-0.2]	0.1 [<0.1-0.2]	0.1 [<0.1-0.2]
Angola	2.4 [1.7-3.2]	2.1 [1.4-2.9]	1.9 [1.3-2.6]	1.8 [1.2-2.4]
Argentina	No data	No data	No data	No data
Armenia	0.2 [0.1-0.3]	0.1 [0.1-0.2]	0.1 [0.1-0.2]	0.1 [<0.1-0.2]
Australia	0.2 [0.1-0.2]	0.1 [0.1-0.2]	0.1 [0.1-0.2]	0.1 [0.1-0.2]
Austria	No data	No data	No data	No data
Azerbaijan	0.2 [0.1-0.2]	0.1 [0.1-0.2]	0.1 [0.1-0.1]	<0.1 [<0.1-0.1]
Bahamas	3.2 [3.1-3.5]	3.4 [3.4-3.5]	3.4 [3.4-3.5]	3.5 [3.4-3.5]
Bangladesh	<0.1 [<0.1-0.1]	<0.1 [<0.1-<0.1]	<0.1 [<0.1-<0.1]	<0.1 [<0.1-<0.1]
Barbados	0.9 [0.7-1.2]	0.9 [0.7-1.2]	0.8 [0.7-1.1]	0.8 [0.6-1.0]
Belarus	0.5 [0.5-0.5]	0.3 [0.3-0.4]	0.2 [0.2-0.3]	0.1 [0.1-0.2]
Belgium	No data	No data	No data	No data
Belize	1.5 [1.3-1.7]	1.6 [1.5-1.7]	1.7 [1.6-1.8]	1.8 [1.5-2.2]
Benin	1.1 [1.1-1.2]	1.2 [1.1-1.3]	1.3 [1.2-1.4]	1.5 [1.4-1.7]
Bhutan	0.1 [0.1-0.4]	0.1 [0.1-0.2]	0.1 [0.1-0.1]	<0.1 [<0.1-0.1]
Bolivia (Plurinational State of)	0.2 [0.1-0.4]	0.3 [0.2-0.5]	0.4 [0.3-0.6]	0.6 [0.4-0.8]
Botswana	21.9 [20.8-23.1]	23.6 [22.5-24.9]	25.4 [24.3-26.7]	27.7 [26.5-29.0]
Brazil	0.5 [0.5-0.6]	No data	No data	No data
Bulgaria	No data	No data	No data	No data
Burkina Faso	0.9 [0.8-1.1]	1.0 [0.9-1.2]	1.3 [1.2-1.5]	2.2 [1.8-2.5]
Burundi	1.0 [0.9-1.1]	1.5 [1.3-1.6]	2.1 [1.9-2.4]	2.5 [2.2-2.7]
Côte d'Ivoire	[illegible]	[illegible]	[illegible]	[illegible]

Prevalence dataset before cleaning

prevalence_Country	prevalence_2013	prevalence_CI_2013	prevalence_2009	prevalence_CI_2009	prevalence_2005	prevalence_CI_2005	prevalence_2001	prevalence_CI_2001
Afghanistan	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1
Albania	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1
Algeria	0.1	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.2
Angola	2.4	1.7-3.2	2.1	1.4-2.9	1.9	1.3-2.6	1.8	1.2-2.4
Argentina	Nodata	NA	Nodata	NA	Nodata	NA	Nodata	NA
Armenia	0.2	0.1-0.3	0.1	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.2
Australia	0.2	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.2
Austria	Nodata	NA	Nodata	NA	Nodata	NA	Nodata	NA
Azerbaijan	0.2	0.1-0.2	0.1	0.1-0.2	0.1	0.1-0.1	0.1	0.1-0.1
Bahamas	3.2	3.1-3.5	3.4	3.4-3.5	3.4	3.4-3.5	3.5	3.4-3.5
Bangladesh	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1	0.1	0.1-0.1
Barbados	0.9	0.7-1.2	0.9	0.7-1.2	0.8	0.7-1.1	0.8	0.6-1.0
Belarus	0.5	0.5-0.5	0.3	0.3-0.4	0.2	0.2-0.3	0.1	0.1-0.2
Belgium	Nodata	NA	Nodata	NA	Nodata	NA	Nodata	NA
Belize	1.5	1.3-1.7	1.6	1.5-1.7	1.7	1.6-1.8	1.8	1.5-2.2
Benin	1.1	1.1-1.2	1.2	1.1-1.3	1.3	1.2-1.4	1.5	1.4-1.7
Bhutan	0.1	0.1-0.4	0.1	0.1-0.2	0.1	0.1-0.1	0.1	0.1-0.1
Bolivia (Plurinational State of)	0.2	0.1-0.4	0.3	0.2-0.5	0.4	0.3-0.6	0.6	0.4-0.8
Botswana	21.9	20.8-23.1	23.6	22.5-24.9	25.4	24.3-26.7	27.7	26.5-29.0
Brazil	0.5	0.5-0.6	Nodata	NA	Nodata	NA	Nodata	NA
Bulgaria	Nodata	NA	Nodata	NA	Nodata	NA	Nodata	NA
Burkina Faso	0.9	0.8-1.1	1.0	0.9-1.2	1.3	1.2-1.5	2.2	1.8-2.5
Burundi	1.0	0.9-1.1	1.5	1.3-1.6	2.1	1.9-2.4	2.5	2.2-2.7
Côte d'Ivoire	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Prevalence Dataset after cleaning

CI_2009	prevalence_2005	prevalence_CI_2005	prevalence_2001	prevalence_CI_2001	Deaths_2013	Deaths_CI_2013	Deaths_2009	Deaths_CI_2009	Deaths_2005	Deaths_CI_2005	Deaths_2001	Deaths_CI_2001	numberInf_2013	numberInf_CI_2013	numberInf_2009	numberInf_CI_2009	n
	0.1	0.1-0.1	0.1	0.1-0.1	500	200-1100	500	200-1000	200	100-500	200	100-500	4500	1700-17000	3400	1500-12000	2
	0.1	0.1-0.1	0.1	0.1-0.1	100	100-100	100	100-100	100	100-100	100	100-100	1000	500-1100	500	500-1000	9
	0.1	0.1-0.2	0.1	0.1-0.2	1400	1000-3300	1100	500-3500	1000	500-3700	1000	500-3000	25000	13000-43000	20000	9300-43000	1
	1.9	1.3-2.6	1.8	1.2-2.4	12000	6300-18000	9800	5300-15000	11000	7200-15000	7800	3300-12000	250000	180000-340000	200000	140000-270000	1
	0.1	0.1-0.2	0.1	0.1-0.2	200	200-500	200	100-500	200	100-500	100	100-200	3700	2400-5500	2300	1800-4600	2
	0.1	0.1-0.2	0.1	0.1-0.2	100	100-200	100	100-100	100	100-200	500	200-500	28000	26000-34000	25000	23000-30000	2
	0.1	0.1-0.1	0.1	0.1-0.1	1000	500-1000	500	500-1000	500	200-500	100	100-200	9200	6700-12000	6900	4900-9500	4
	3.4	3.4-3.5	3.5	3.4-3.5	1000	1000-1000	500	500-500	500	500-500	500	500-1000	7700	7300-8300	7600	7300-7900	6
	0.1	0.1-0.1	0.1	0.1-0.1	500	200-3000	500	200-1000	500	100-500	100	100-500	9500	4100-97000	7000	3300-30000	9
	0.8	0.7-1.1	0.8	0.6-1.0	100	100-100	100	100-100	100	100-100	100	100-200	1700	1300-2200	1600	1200-2000	1
	0.2	0.2-0.3	0.1	0.1-0.2	1000	1000-1200	1000	1000-1000	1000	500-1000	500	200-500	25000	24000-27000	19000	17000-21000	1
	1.7	1.6-1.8	1.8	1.5-2.2	200	100-200	200	200-200	200	200-200	100	100-200	3300	2900-3600	3000	2800-3300	2
	1.3	1.2-1.4	1.5	1.4-1.7	2700	2300-3200	3100	2700-3600	4800	4400-5400	4100	3600-4600	74000	69000-80000	67000	63000-73000	8
	0.1	0.1-0.1	0.1	0.1-0.1	100	100-100	100	100-100	100	100-100	100	100-100	1000	500-2100	500	500-1000	9
	0.4	0.3-0.6	0.6	0.4-0.8	1200	1000-2900	1000	1300-4000	2200	1600-5300	2100	1400-5600	15000	7900-33000	13000	12000-36000	2
	25.4	24.3-26.7	27.7	26.5-29.0	5800	5000-6500	8200	7200-9700	14000	13000-16000	20000	19000-22000	320000	310000-340000	310000	290000-320000	2
	1.3	1.2-1.5	2.2	1.8-2.5	5800	4600-7300	8000	6500-9900	14000	12000-17000	19000	16000-23000	110000	100000-130000	120000	100000-130000	1
	2.1	1.9-2.4	2.5	2.2-2.7	4700	3900-5600	7200	6200-8500	7900	6800-9200	5700	4600-7100	83000	76000-91000	97000	87000-110000	1
	5.1	4.7-5.5	6.6	5.9-7.3	28000	25000-32000	37000	33000-43000	53000	46000-60000	51000	44000-61000	370000	330000-410000	440000	410000-480000	9
	0.4	0.4-0.5	0.5	0.4-0.6	100	100-100	100	100-100	100	100-200	100	100-200	1500	1300-1800	1400	1200-1700	1
	1.2	0.4-2.4	1.6	0.6-2.9	2200	1000-4000	3900	2000-9300	7600	1900-16000	6200	2300-10000	75000	41000-130000	82000	41000-150000	9
	5.2	4.9-5.5	5.3	4.9-5.7	44000	40000-46000	46000	42000-52000	48000	44000-54000	38000	34000-44000	600000	560000-650000	600000	560000-640000	9
	6.8	5.9-7.8	8.7	7.2-10.6	11000	9500-12000	11000	9300-13000	16000	13000-20000	15000	11000-20000	120000	110000-130000	140000	120000-150000	1
	3.6	3.1-4.3	3.5	3.0-4.2	15000	12000-18000	13000	11000-16000	16000	13000-19000	11000	8700-13000	210000	170000-250000	220000	190000-260000	2
	0.3	0.2-0.5	0.4	0.2-0.5	1000	200-1600	1300	500-2600	1700	500-3100	2100	1000-3300	38000	23000-59000	33000	20000-48000	3
	3.9	3.6-4.2	5.0	4.6-5.6	5400	4900-6000	6900	6200-7600	9300	8400-10000	9600	8500-11000	69000	64000-75000	77000	71000-82000	8
	0.2	0.1-0.2	0.2	0.2-0.3	500	200-500	200	100-500	200	200-500	200	100-200	7600	5400-9100	7100	4800-8500	9

Sample of merged dataset (NA omitted)

Creating a NoSQL database

A columnar type NoSQL database (mongoDB) is used to store the two retrieved data sets and we perform a set of queries on the databases to see that all the data is loaded correctly. Two data bases are used for storage as the two data sets are of a completely different purposes but have the same aim. So the data retrieved from the two databases simultaneously can give insight to new results as one has the medical supply data and the other has the statistical facts.

Code:

```
mongoData<-mongo("SCMS")

mongoData$insert(supplyData)

mongoData1<-mongo("HIVdata")

mongoData1$insert(finalData)
```

Queries:

```
> mongoData$count('{"Country":"Vietnam"}')
[1] 2064
> yeard<-mongoData$find('{"Year":2008}')
 Imported 2216 records. Simplifying into dataframe...
> dat <- mongoData$find('{"Product_Name":"Nevirapine"}', fields = '{"_id":0,"Pack_Price":1, "Country":1,"Year":1}')
 Imported 1698 records. Simplifying into dataframe...
> countryd<-mongoData1$find('{"Country":"Afghanistan"}')
 Imported 2 records. Simplifying into dataframe...
> mxdel<-mongoData$aggregate('[{"$group":{"_id":"$Country", "count": {"$sum":1},"max":{"$max":"$Delivery_Time"}}}]')
 Imported 45 records. Simplifying into dataframe...
> View(yeard)
```

Queries with head data:

```
> mongoData$count('{"Country":"Vietnam"}')
[1] 2064
> yeard<-mongoData$find('{"Year":2008}')
Imported 2216 records. Simplifying into dataframe...
> head(yeard)
   ID Project_Code       PO_SO     ASN_DN      Country Managed_By Fulfill_Via Shipment_Mode PO_Sent_to_Vendor_Date Scheduled_Delivery_Date Delivered_t_Client_Date Year
1 262    116-ZA-T01 SCMS-14050 ASN-1251 South Africa   PMO - US Direct Drop           N/A    2007-11-11 19:00:00     2008-01-28 19:00:00     2008-01-28 19:00:00 2008
2 269    108-VN-T01 SCMS-14190 ASN-1192      Vietnam   PMO - US Direct Drop           Air    2007-11-18 19:00:00     2008-01-20 19:00:00     2008-01-20 19:00:00 2008
3 270    108-VN-T01 SCMS-14200 ASN-1171      Vietnam   PMO - US Direct Drop           Air    2007-11-20 19:00:00     2008-01-20 19:00:00     2008-01-20 19:00:00 2008
4 284    107-RW-T01 SCMS-14630 ASN-1253       Rwanda   PMO - US Direct Drop           Air    2007-12-09 19:00:00     2008-01-30 19:00:00     2008-01-30 19:00:00 2008
5 305    123-NG-T01 SCMS-15550 ASN-1269      Nigeria   PMO - US Direct Drop           Air                   <NA>     2008-02-04 19:00:00     2008-02-04 19:00:00 2008
6 343    116-ZA-T01 SCMS-16600 ASN-1520 South Africa   PMO - US Direct Drop           N/A    2008-01-03 19:00:00     2008-01-20 19:00:00     2008-01-03 19:00:00 2008
  Delivery_Time Delay_Time Product_Group                                                     Vendor                       Molecule_Test_Type
1            78          0           ARV                                    JSI R&T INSTITUTE, INC.                               Zidovudine
2            63          0           ARV                             GILEAD SCIENCES IRELAND, INC.          Tenofovir Disoproxil Fumarate
3            61          0           ARV SUN PHARMACEUTICAL INDUSTRIES LTD (RANBAXY LABORATORIES LIMITED)                          Zidovudine
4            52          0          HRDT                                   Orgenics, Ltd HIV, Lancet, Safety, for HIV Test kits, 100 Pcs
5             0          0          HRDT                                   Orgenics, Ltd HIV, Lancet, Safety, for HIV Test kits, 100 Pcs
6             0        -17           ARV                                    JSI R&T INSTITUTE, INC.                                Stavudine
                                   Product_Name          Molecule_First_Test Dosage_ml_mg Dosage_Unit_mg              Dosage_Form Unit_of_Measure_Per_Pack
1                                     Retrovir                   Zidovudine           10             10            Oral solution                      200
2                                       Viread Tenofovir Disoproxil Fumarate        <NA>            300                   Tablet                       30
3                                   Zidovudine                   Zidovudine         <NA>            300                   Tablet                       60
4 HIV, Lancet, Safety, for HIV Test kits, 100 Pcs                        test         <NA>           <NA>    Test kit - Ancillary                      100
5 HIV, Lancet, Safety, for HIV Test kits, 100 Pcs                        test         <NA>           <NA>    Test kit - Ancillary                      100
6                                        Zerit                    Stavudine            1              1 Powder for oral solution                      200
  Line_Item_Quantity Pack_Price Donation_Designation First_Line_Designation Weight_Kilograms Line_Item_Insurance Dosage_First Dosage_Second Dosage_Third Dosage_mg
1                 40       7.94                    N                    Yes                0                0.51           10             0            0      <NA>
2               1000      17.00                    N                    Yes               76               27.20          300             0            0       300
3                400       8.38                    N                    Yes               23                5.38          300             0            0       300
4               1500       0.01                    Y                    Yes               99                0.02            0             0            0      <NA>
5                650       0.01                    Y                    Yes                0                0.01            0             0            0      <NA>
6                200       1.70                    Y                    Yes                0                0.54            1             0            0      <NA>
  Molecule_Second Molecule_Third
1            <NA>           <NA>
2            <NA>           <NA>
3            <NA>           <NA>
4            <NA>           <NA>
5            <NA>           <NA>
6            <NA>           <NA>
```

Queries being run to test the database

Continuation of test Queries

```
> dat <- mongoData$find('{"Product_Name":"Nevirapine"}', fields = '{"_id":0,"Pack_Price":1, "Country":1,"Year":1}')
 Imported 1698 records. Simplifying into dataframe...
> head(dat)
       Country Year Pack_Price
1      Vietnam 2006       6.20
2     Tanzania 2006       3.65
3      Vietnam 2007       1.92
4      Vietnam 2007       1.92
5      Vietnam 2008       2.90
6 South Africa 2008       4.90
> countryd<-mongoData1$find('{"Country":"Afghanistan"}')
 Imported 2 records. Simplifying into dataframe...
> head(countryd)
      Country prevalence_2013 prevalence_CI_2013 prevalence_2009 prevalence_CI_2009 prevalence_2005 prevalence_CI_2005 prevalence_2001 prevalence_CI_2001 Deaths_2013
1 Afghanistan             0.1            0.1-0.1             0.1            0.1-0.1             0.1            0.1-0.1             0.1            0.1-0.1         500
2 Afghanistan             0.1            0.1-0.1             0.1            0.1-0.1             0.1            0.1-0.1             0.1            0.1-0.1         500
  Deaths_CI_2013 Deaths_2009 Deaths_CI_2009 Deaths_2005 Deaths_CI_2005 Deaths_2001 Deaths_CI_2001 numberInf_2013 numberInf_CI_2013 numberInf_2009 numberInf_CI_2009
1       200-1100         500       200-1000         200        100-500         200        100-500           4500        1700-17000           3400        1500-12000
2       200-1100         500       200-1000         200        100-500         200        100-500           4500        1700-17000           3400        1500-12000
  numberInf_2005 numberInf_CI_2005 numberInf_2001 numberInf_CI_2001
1           2500         1100-9200           1700         1000-5200
2           2500         1100-9200           1700         1000-5200
> mxdel<-mongoData$aggregate('[{"$group":{"_id":"$Country", "count": {"$sum":1},"max":{"$max":"$Delivery_Time"}}}]')
 Imported 45 records. Simplifying into dataframe...
> head(mxdel)
            _id count max
1        Angola    21 183
2       Nigeria  3582 341
3      Tanzania  1557 384
4 Côte d'Ivoire  1083  NA
5         Haiti  1965 545
6       Vietnam  2064 315
> |
```

Problems Faced:

The main problems that I faced is when I tried to combine the two data sets that I retrieved. I have tried very hard to somehow embed into each other, but then I could only complete the process till the merging of the three datasets retrieved from WHO data. I think the main problem lies in the fact that a nested loop is necessary to embed the HIV statistics in to the supply chain data by inspecting each element and then merging them. The process took a lot of time and I finally abandoned that idea to create two separate datasets

Conclusion:

- Hence we can see that the supply chain data which is available publicly can be used for conducting effecting analytics if proper preprocessing is performed on the data.
- We also realize the supply chain data can also be linked to various factors in a given environment.
- If properly realized, the HIV medical supplies supply chain network can be properly optimized to the HIV prevalence rate of a country, so that the people who need help the most will receive it. But this cannot be achieve without further improvement in the data cleaning and acquisition systems.

References:

a. https://cran.r-project.org
b. http://scms.pfscm.org/scms/about
c. http://www.pepfar.gov/
d. http://www.who.int/en/
e. Class Notes 'Collecting, Storing and Retrieving Data'-Yatish Jain and Martin Schedlbauer