Donald Dunagan

26 November 2018

Documentation of ARDI Data Preparation

The Alcohol-Related Disease Impact (ARDI) application is a tool that aids public health professionals in evaluating the health impacts of alcohol. The ARDI application was created by the Centers for Disease Control and Prevention at the national level, but with data able to be viewed by state. Measured impacts include alcohol-attributable deaths (AAD), or, estimated total number of deaths attributable to alcohol, years of potential life lost (YPLL), or, estimated total number of alcohol-related years of life lost from premature death, and alcohol-attributable fractions (AAF), or, estimated proportions of deaths from different causes that are attributable to alcohol. The data undergoing preparation in this document is specifically the ARDI data collected from 2006 to 2010 for the state of Georgia.

The first step in the ARDI data preparation process was to pull all of the raw data, three separate Excel spreadsheets, into Qlik Sense (QS). This was done with three QS web file data connections. The tier 1 load script took the tables AAF, AlcoholAttrDeaths, and YPLL from the data connections DBHDD - ARDI Alcohol-Attributable Fractions, DBHDD - ARDI Alcohol-Attributable Deaths, and DBHDD - ARDI Years of Potential Life Lost, respectively, removed 4 rows of excess header, and saved the data as three separate .QVD files: Alcohol\_Attributable\_Fractions, Alcohol\_Attributable\_Deaths, and Years\_Life\_Lost.

The tier 2 load script would involve taking the tier 1 .QVD files, transforming the data into the final data model, and saving the data model as .QVD files. Essentially, I did this process twice: the first time implementing my own idea for the final data model and the second, implementing Hajime’s. The two models are a bit different and even though Hajime’s model was the one ultimately used, I will describe them both.

The following documents my initial attempt. The Alcohol\_Attributable\_Deaths and Years\_Life\_Lost tables had similar structure and were straightforward to format. The values for [Harmful Effect] were extracted from the first column with either ‘Chronic’ or ‘Acute’ being hardcoded for [Harmful Effect Type] depending on where the harmful effect occurred in the table as the acute effects were separate and came after the chronic effects. The column containing overall measure was dropped as aggregation between genders is simple enough in the QS front end and the columns containing data for males and females were pivoted resulting in a single [Gender] column. The measure for Alcohol\_Attributable\_Deaths was [#Deaths] while the measure for Years\_Life\_Lost was [#YPLL]. Because Alcohol\_Attributable\_Deaths and Years\_Life\_Lost are structured similarly aside from their measure, they were combined together resulting in Alcohol\_Attributable\_Deaths\_Years\_Life\_Lost.

The Alcohol\_Attributable\_Fractions data was much more difficult to wrangle because the data were arranged as a report rather than a table. The report would alternate between direct alcohol-attributable fractions which has a single measure (direct alcohol-attributable fraction) and indirect alcohol-attributable fractions which has two measures (medium and high indirect alcohol attributable fraction). Additionally, there was a column for age buckets, but only one harmful effect, motor-vehicle traffic crashes had this data. Another troubling aspect was that, in the first column, the report would give the harmful effect and then have ‘Males’ and ‘Females’ rows immediately underneath it for the stats.

My initial approach for tackling this was to count the beginning and end rows for each alternating ‘Direct’ and ‘Indirect’ fractions section of the report, hardcode in [Harmful Effect Type], [Attributable Fraction Type], and [Gender] and only pull the attributable fraction data and in the case of the motor-vehicle traffic crashes, the age data. This was all done in the tier 2 Alcohol\_Attributable\_Fractions table. This brute force method was not efficient and certainly not a great way to approach this problem, but it did work.

The second model, the one that Hajime proposed, was much more elegant. The idea was to construct a series of fact tables and then link them together by what they had in common. Extracting the necessary data from Alcohol\_Attributable\_Deaths and Years\_Life\_Lost was almost identical to how I did it the first time. However, the first time, I manually counted the beginning and end rows to extract as chronic harmful effects and acute harmful effects. At the recommendation of Hajime, I refactored the script to incorporate variables which exactly identify the row numbers where the data I need begin and end. While this dataset is fairly small, I see that utilizing variables is a smarter approach as it can help prevent human error and is scalable to larger datasets. The structured data were stored in the tables FACT\_AAD and FACT\_YPLL.

Next was working with Alcohol\_Attributable\_Fractions. To create the FACT\_DAAF table, [Harmful Effect], [Harmful Effect Type], [Gender], and [#Direct AAF] were pulled from Alcohol\_Attributable\_Fractions for the direct alcohol-attributable fractions, excluding data related to motor vehicle traffic crashes. To create the IAAF table, [Harmful Effect] [Gender], and [#Indirect AAF] were pulled from Alcohol\_Attributable\_Fractions for the indirect alcohol-attributable fractions with a pivot for [Attribute Degree] and ‘Chronic’ hardcoded as [Harmful Effect Type]. To create the MVTC table, [Gender], [Age Group] and [#MVTC Direct AAF] were extracted from Alcohol\_Attributable\_Fractions for the motor vehicle traffic crashes fractions, hardcoding ‘Acute’ as [Harmful Effect Type] and ‘Motor-vehicle traffic crashes’ as [Harmful Effect].

With the data extracted, the final step was to structure the data model. FACT\_AAD, FACT\_YPLL, and FACT\_DAAF were joined together into FACT\_AAD\_YPLL\_DAAF as they were all very similar in structure. Then, [Harmful Effect Type], [Harmful Effect], and [Gender] were hashed for FACT\_AAD\_YPLL\_DAAF, IAAF, and MVTC resulting in a unique hash string for each combination of the features across the tables. LNK\_Effect\_EffectType\_Gender was created as the table to link together all of the fact tables and included [Harmful Effect Type], [Harmful Effect], and [Gender] and the corresponding [%Effect\_EffectType\_Gender] hashes. Lastly, [Harmful Effect Type], [Harmful Effect], and [Gender] were dropped from all of the fact tables as because of the linking hash, they were no longer needed.

As a personal reflection, I certainly learned a lot while working on these data load scripts: from QS scripting, to data cleaning and restructuring, and data modeling. While the Alcohol\_Attributable\_Fractions data were certainly difficult to work, it was a good challenge that pushed me to grow and learn.