Workflow of Data Exploration

Nibret Daba

University of Maryland

College of Information Studies

October 2015

Table of Contents

[Step 1: Finding Dataset 3](#_Toc432770765)

[Data Source 3](#_Toc432770766)

[Dataset 3](#_Toc432770767)

[Step 2: Formulating Research Question 4](#_Toc432770768)

[Step 3: Data Filtering 4](#_Toc432770769)

[Step 4: Exporting and Cleaning Data 5](#_Toc432770770)

[Step 5: Merging Data 6](#_Toc432770771)

[Step 6: Documentation 10](#_Toc432770772)

**Information Organization Assignment**

This document is a full-workflow of data exploration process. By analyzing through the data, the intention is to find out if issuing more liquor license has anything to do with violence and vehicle accident in Howard County, MD.

## Step 1: Finding Dataset

### Data Source

Three datasets are identified from three different data sources.

1. Howard County Open Data Portal: (<https://opendata.howardcountymd.gov/>)
2. State of Maryland Open Data Portal: (<https://data.maryland.gov/>)
3. Comptroller of Maryland Online Database: (<https://interactive.marylandtaxes.com/webapps/licprt/user/ilu_QueryRetailer.asp>)

### Dataset

Howard County Data Portal, Howard County Maryland. (2015). *Howard County Police Department Call for Service: 2014*. Retrieved from <https://opendata.howardcountymd.gov/Public-Safety/Howard-County-Police-Department-Call-For-Service-2/qccx-65fg> on 2015, September 29.

This dataset contains records of calls that are made to request police service in the year 2014. It is located under Public Safety category of the portal and is available for public use. The dataset contains 130,538 records. The records comprise reason of call, date, location, time, statistical reporting area and beat information.

Maryland Government, Department of Information Technology. (2013). *2012 Vehicle Collisions Investigated by State Police*. Retrieved from <https://data.maryland.gov/Public-Safety/2012-Vehicle-Collisions-Investigated-by-State-Poli/pdvh-tf2u> on 2015, September 23.

This dataset contains records of vehicle collisions investigated by State Police in the year 2012. It has accident date and time, location, county, days of the week, road among other things and it is provided by Maryland State Police.

Comptroller of Maryland. (2015). *Alcohol and Tobacco Tax, Alcoholic Beverage Retail Licenses* [Database Record]*.* Retrieved from Alcoholic Beverage Retail Licenses Database. <https://interactive.marylandtaxes.com/webapps/licprt/user/ilu_QueryRetailer.asp> on 2015, October 04.

This dataset contains records of alcohol and tobacco tax payers’ information with active licenses. The result from the database query contains tax registration number, county, date, corporate and trade names of the businesses.

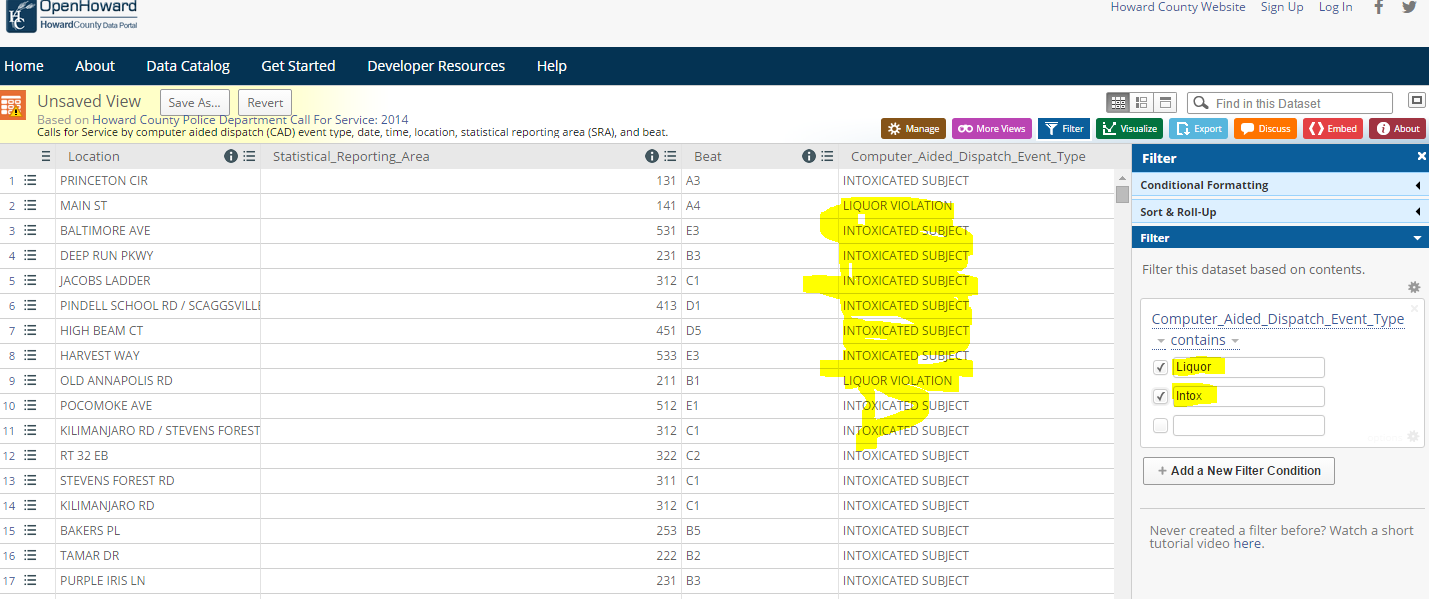
## Step 2: Formulating Research Question

Based on the datasets at hand, the proposed research question would be “Does issuing more liquor license aggravate violence and vehicle accident in Howard County MD?”

## Step 3: Data Filtering

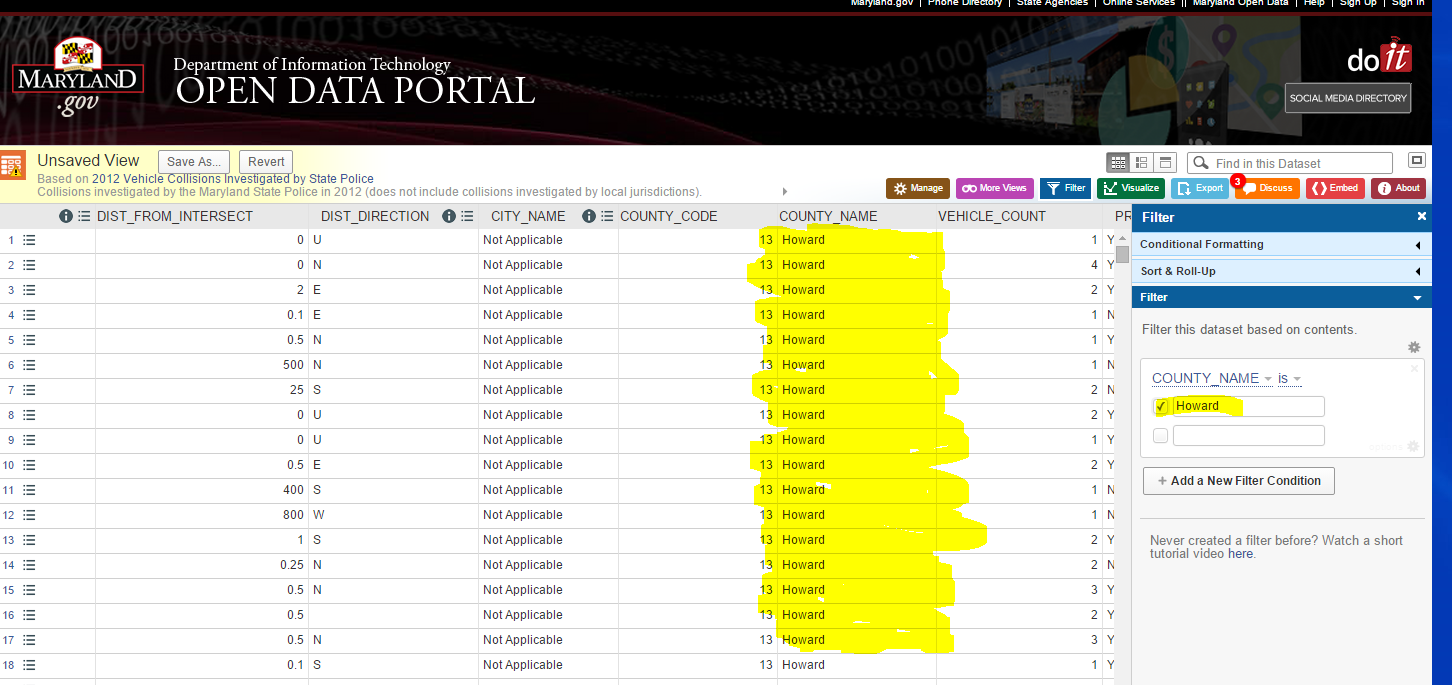
**3.1.** Dataset 1: Get records “**where reason of call is** **Liquor Violation and Intoxicated Subject.”**

**Technique:** The **“Filter”** feature from the portal is used to complete this task. Below is the screen shot showing the filters used and the result (Highlighted Yellow)



**3.2.** Dataset 2: Get records “**where collision happened only in Howard County, MD”**.

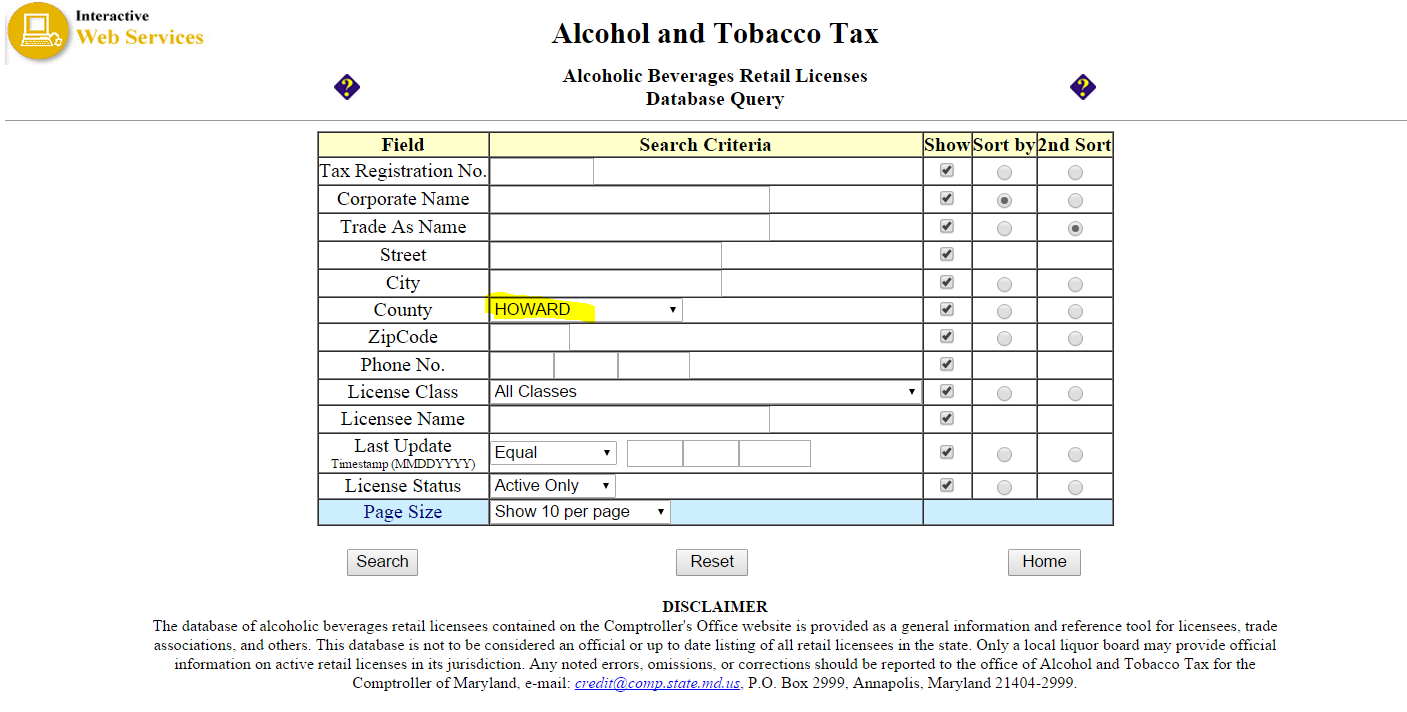
**Technique:** “Filter” feature of the portal is used. Below is the screen shot.



**3.3**. Dataset 3: Get records **“where alcohol and tobacco businesses located at Howard County, MD”.**

**Technique:**  While querying the database select “Howard” from the County drop down list.

Screen shot:

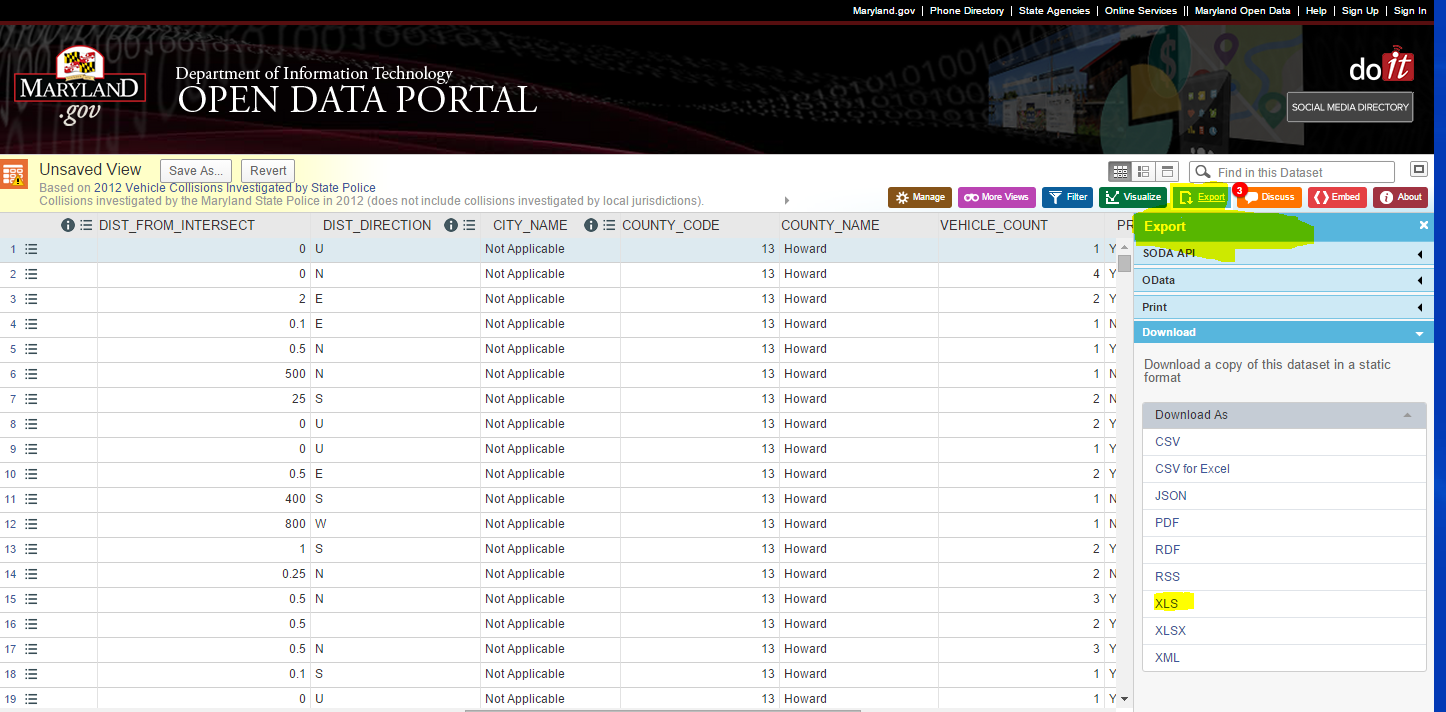


## Step 4: Exporting and Cleaning Data

* 1. Exporting to Excel

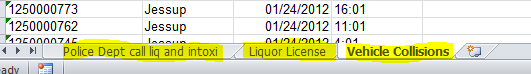
**Techniques:**

* + 1. For dataset one and two, the **“Export”** feature of the data portal is used. Here below is the screen shot.



* + 1. For dataset three, after running the query against the database, the result is copied and pasted to Excel.

Here below is a screen shot where the three datasets sitting on separate Excel tab.



* 1. Cleaning Data

**Target:** Taking out extra spaces and irrelevant characters from the **Location** field of each dataset. This is because **Location** is the common field between the three datasets and is a key field to merge on.

**Technique:** Excel **Trim and LEFT** functions is used.

**Result:** The three datasets have common column with the same data structure to join.

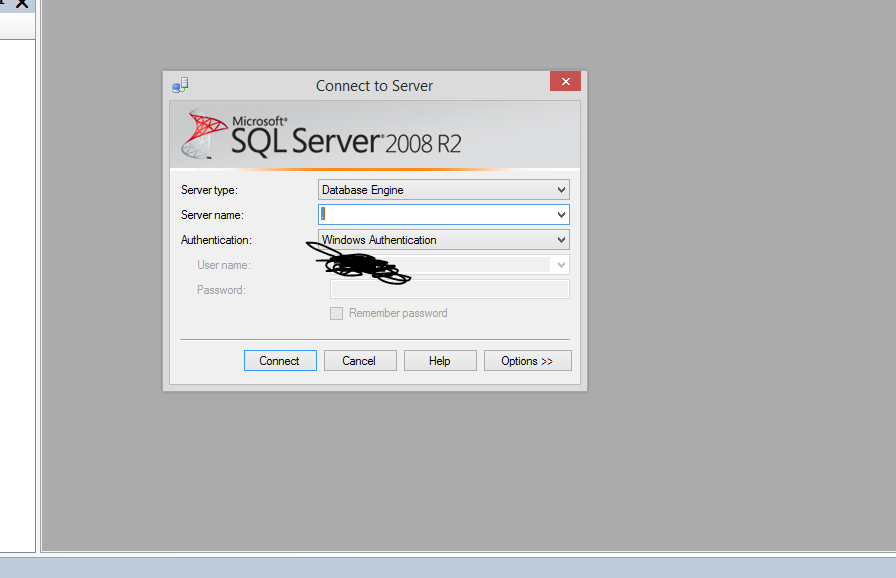
## Step 5: Merging Data

The merging between the three data sets is accomplished using SQL script. For that, the datasets are imported to SQL Server into three separate tables. The following steps and screen shots illustrate how the datasets get into SQL Server.

* 1. Importing from Excel to SQL Server

**Techniques:**

* + 1. Connecting to SQL Server

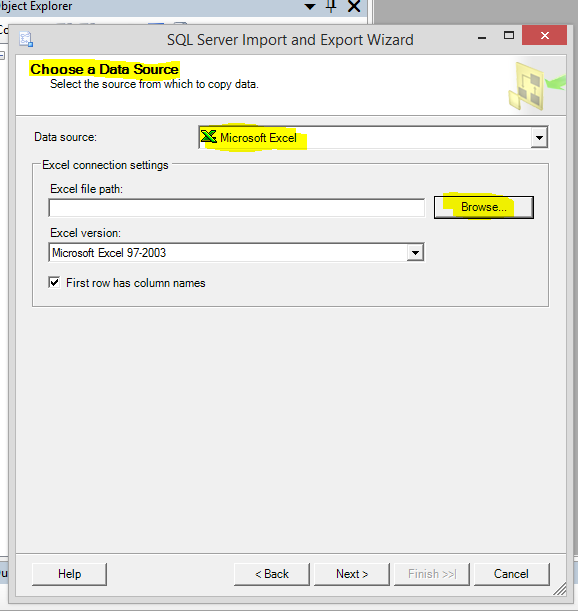


* + 1. Creating a database and importing the excel files

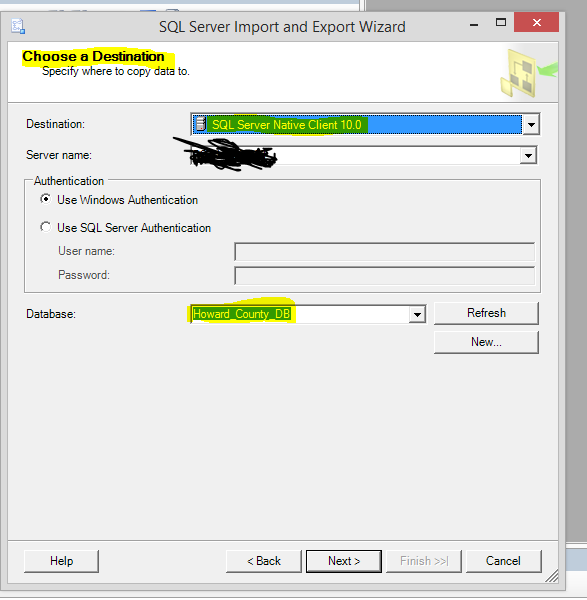
Database Creating Script:

Create database [Howard\_County\_DB]

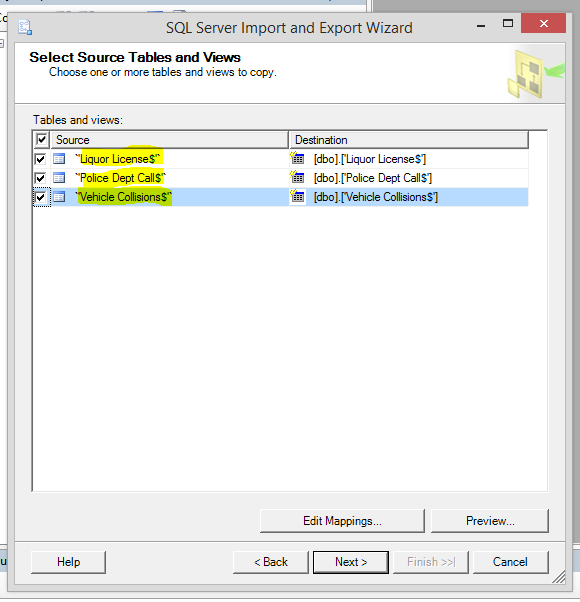
* + 1. Importing the datasets to the newly created database
* Right click on to the database ([Howard\_County\_DB]>>>Task>>>Import Data
* Pass the welcome page
* Choose Microsoft Excel as a data source
* Browse to the excel file

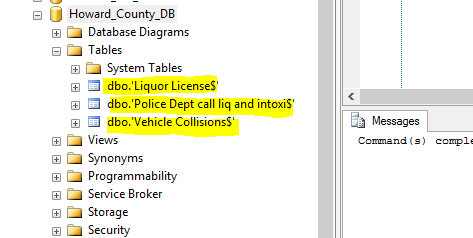


* Choose SQL Server as a destination



* Choose the three data sets to import



* The three datasets are imported to SQL Server table
* 
  1. Write a Query to Join the three tables:

use Howard\_County\_DB

Select \*

from dbo.['Liquor License$'] as a

Join dbo.['Police Dept call$'] as b

on a.[StreetAddr1] = b.[Location]

join dbo.['Vehicle Collisions$'] as c

on a.[StreetAddr1] = c.[INTERSECT\_ROAD]

## Step 6: Documentation

1. Write the README file
2. Creating Github account
3. Upload the following documents to a Github repository:
   1. README file
   2. The Processing Document
   3. Final dataset
   4. The original datasets(three datasets)