****

**CIS 4560**

**Term Project Tutorial**

**Instructor: Jongwook woo**

**Date: 12/16/2018**

**Lab Tutorial**

Amy Li

Brian Seto

Phillip Nguyen

Tararath Chea

**Visualizing Youth Risk Behavior**

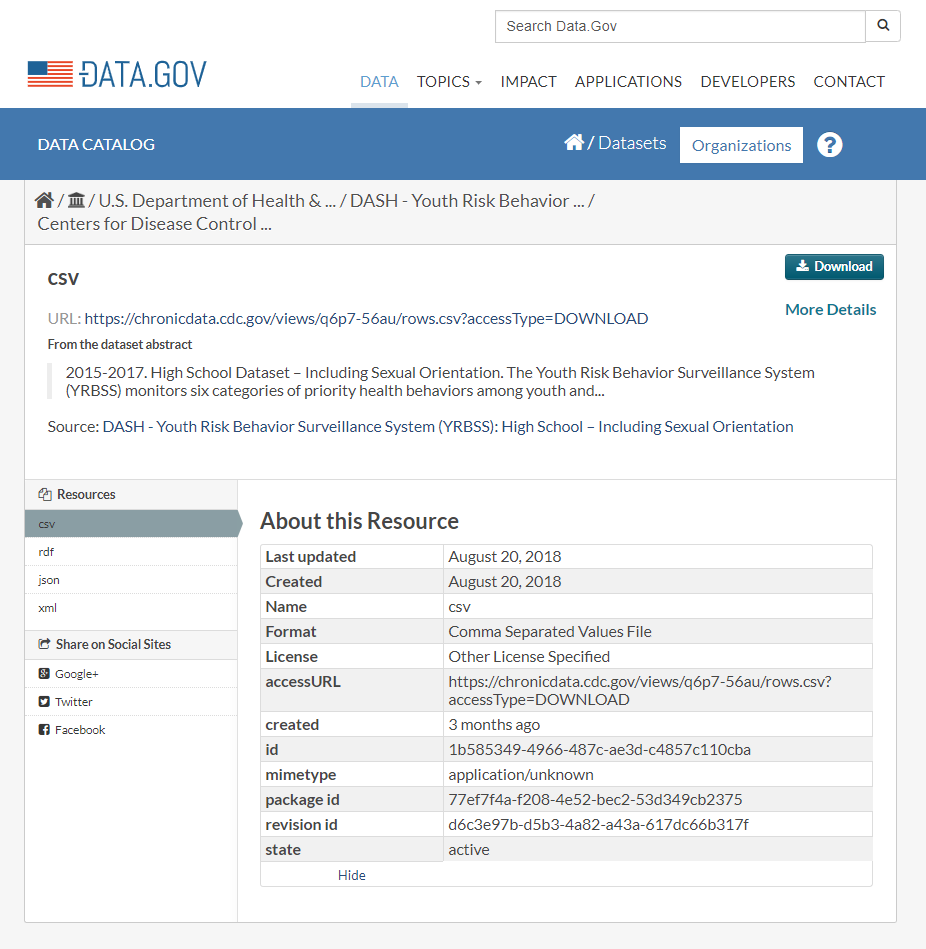
**Objectives**

**In this hands-on lab, you will learn how to:**

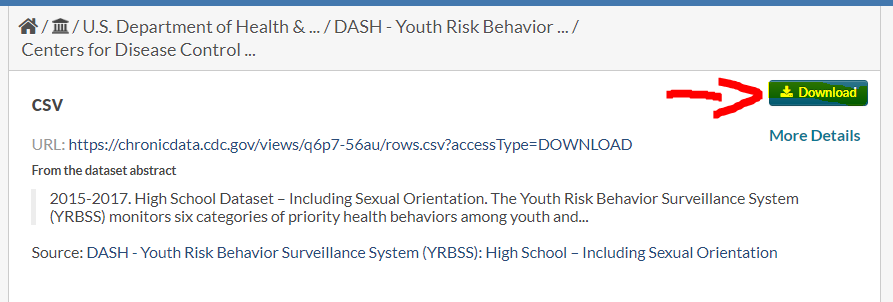
* How to download data from Data.gov
* Input data into Hadoop cluster
* Create a data table in Beeline
* Analyzing table with SQL commands
* Visualization
* Problem Encounters
* GITHUB Tutorial URL: https://github.com/tchea/CIS4560-YouthRisk

**How to get the data from Data.gov**

In order to download the data file, follow the link below to data.gov: <https://catalog.data.gov/dataset/77ef7f4a-f208-4e52-bec2-53d349cb2375/resource/1b585349-4966-487c-ae3d-c4857c110cba>

****

Follow by select ***CSV file*** and click on ***Download***

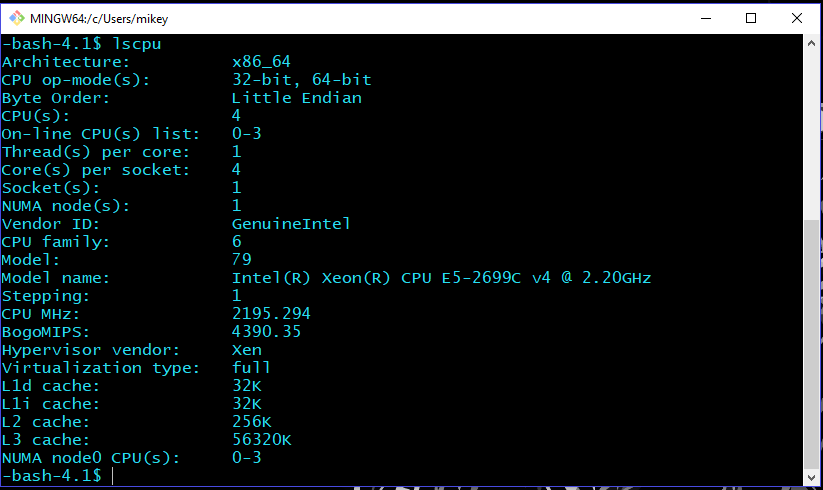
****

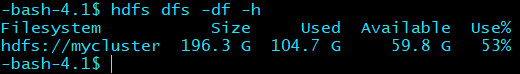
**Platform Spec**

Command to check the CPU spec:

lscpu - to find out the cpu

hdfs dfs -df -h to find the available memory usages





|  |  |
| --- | --- |
| Oracle Compute Edition | 5 Nodes |
| OCPUs | 10 |
| Memory | 150 GB |
| Storage | 678 GB |
| CPU Speed | 2.20 GHz |
| HDFS capacity | 196.3 GB |

**Step 1: Put Dataset into the Server:**

1. Use assigned IP address:

$ ssh (username)@129.150.205.28

1. Try the following HDFS command to see if the file is in the server:

-bash-4.1$ hdfs dfs -ls

1. Download the csv file onto the server:

$ wget -O youthrisk https://chronicdata.cdc.gov/views/q6p7-56au/rows.csv?accessType=DOWNLOAD



1. Make a directory to put the csv file into:

-bash-4.1$ hdfs dfs -mkdir youthrisk

-bash-4.1$ hdfs dfs -put youthrisk youthriskdata

1. Remove the file:

-bash-4.1$ rm youthriskdata

**Step 2: Connect to Beeline**

beeline> !connect jdbc:hive2://cis4560-bdcsce-4.compute-608214094.oraclecloud.internal:2181,cis4560-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis4560-bdcsce-3.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

(Press enter twice)

Use your own database:

Use pnguye47;

***Query for Table Creation***

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> CREATE EXTERNAL TABLE IF NOT EXISTS risk\_data()

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/pnguye47/youthriskdata' TBLPROPERTIES

('skip.header.line.count'='2');

***Table lists according to the actual data field***

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> CREATE EXTERNAL TABLE IF NOT EXISTS risk\_data(YEAR INT, LocationAbbr STRING, LocationDesc STRING, DataSource STRING, Topic STRING, Subtopic STRING, ShortQuestionText STRING, Greater\_Risk\_Question STRING, Description STRING, Data\_Value\_Symbol INT, Data\_Value\_Type STRING, Greater\_Risk\_Data\_Value INT, Greater\_Risk\_Data\_Value\_Footnote\_Symbol INT, Greater\_Risk\_Data\_Value\_Footnote INT, Greater\_Risk\_Low\_Confidence\_Limit INT, Greater\_Risk\_High\_Confidence\_Limit INT, Lesser\_Risk\_Question STRING, Lesser\_Risk\_Data\_Value INT, Lesser\_Risk\_Data\_Value\_Footnote\_Symbol INT, Lesser\_Risk\_Data\_Value\_Footnote INT, Lesser\_Risk\_Low\_Confidence\_Limit INT, Lesser\_Risk\_High\_Confidence\_Limit INT, Sample\_Size INT, Sex STRING, Race STRING, Grade STRING, SexualIdentity STRING, SexOfSexualContacts STRING, GeoLocation INT, TopicId STRING, SubTopicID STRING, QuestionCode STRING, LocationId INT, StratID1 STRING, StratID2 STRING, StratID3 STRING, StratID4 STRING, StratID5 STRING, StratificationType STRING, StratID6 STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

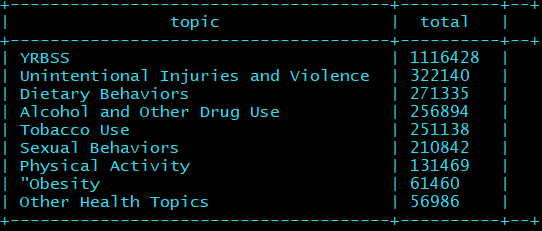
LOCATION '/user/pnguye47/youthriskdata' TBLPROPERTIES

('skip.header.line.count'='2');

***Sample Data Query***

To determine how many risk based on topics are being used:

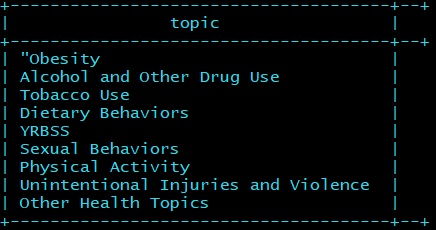
0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT topic, COUNT(topic) as total from risk\_data GROUP BY topic ORDER BY total DESC limit 20;



To find the given data of risk youth:

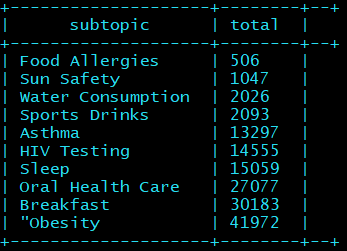
0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT DISTINCT topic FROM risk\_data;

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT LocationDesc FROM risk\_data ORDER BY LocationAbbr DESC limit 10;

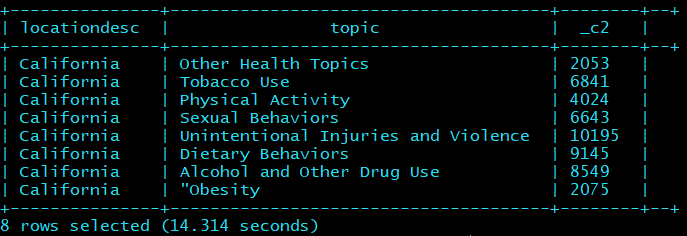


To find the subtopics:

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT subtopic, COUNT(subtopic) as total FROM risk\_data GROUP BY subtopic ORDER BY total ASC limit 10;

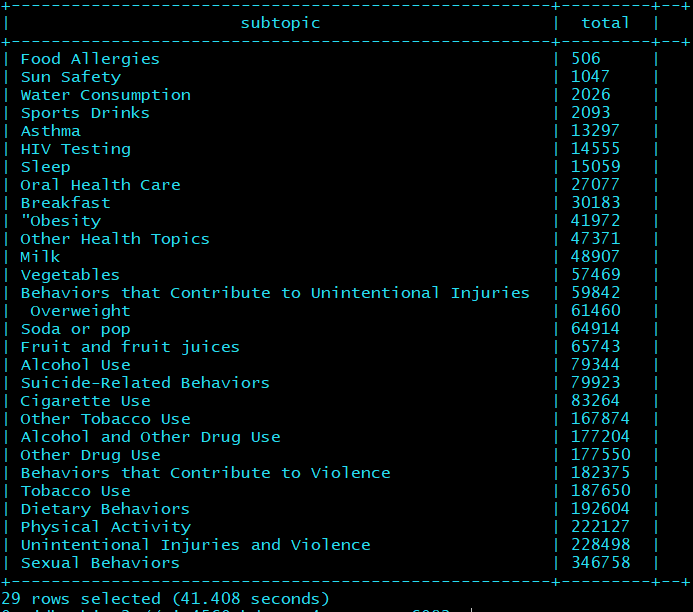


Data Location Specific from California

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082>SELECT locationdesc, topic, count(topic) from risk\_data where locationdesc = 'California' GROUP BY locationdesc, topic;  


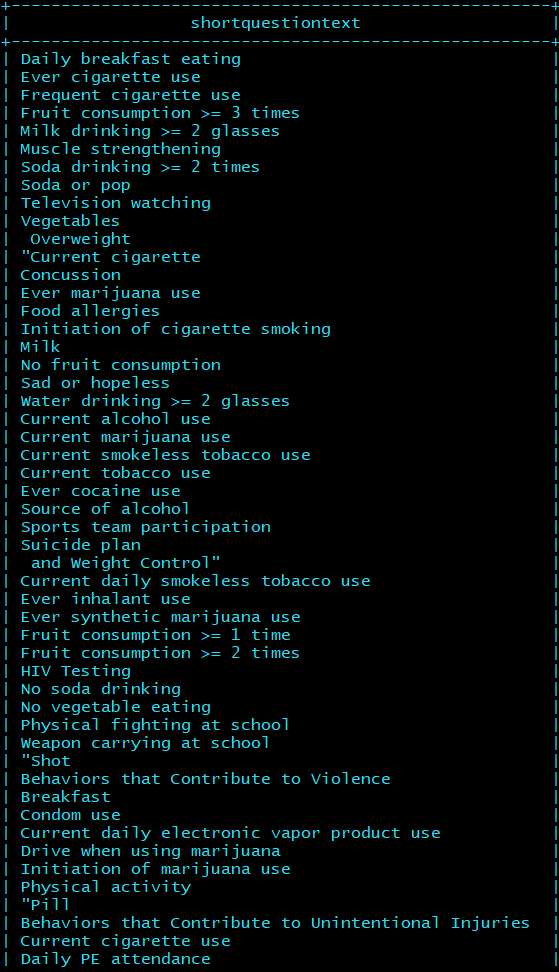
Total Subtopic of each catagory

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT subtopic, COUNT(subtopic) as total FROM risk\_data GROUP BY subtopic ORDER BY total ASC;



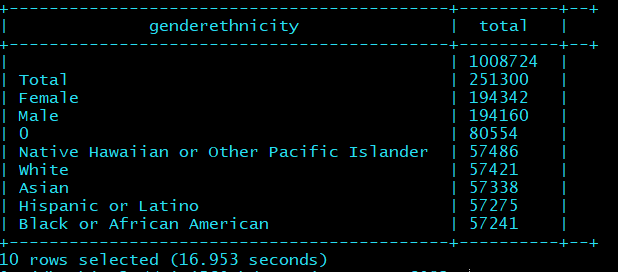
Short Question from the survey

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT DISTINCT shortquestiontext from risk\_data;



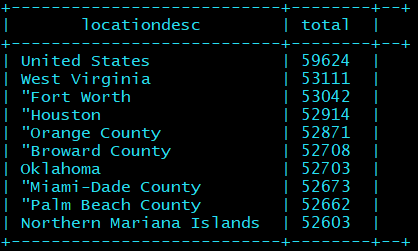
To determine the demographics of the people that are at risk of these:

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT race, COUNT(race) as total FROM risk\_data GROUP BY race ORDER BY total DESC limit 10;



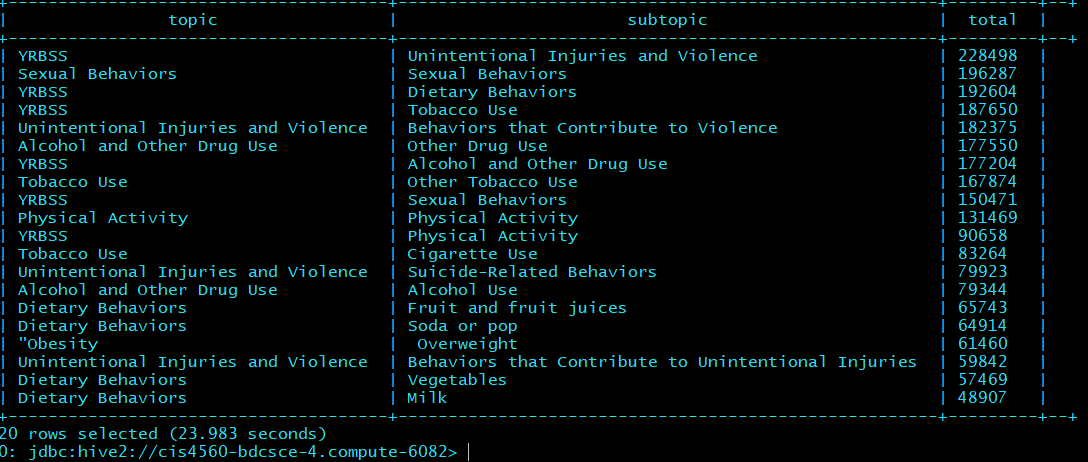
To determine the locations that has been taken down by the dataset:

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT locationdesc, COUNT(locationdesc) as total from risk\_data GROUP BY locationdesc ORDER BY total DESC limit 10;



To determine the total of topic base on subtopic

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT Topic, Subtopic, COUNT(Subtopic) as total FROM risk\_data GROUP BY Topic, Subtopic ORDER BY total DESC limit 20;



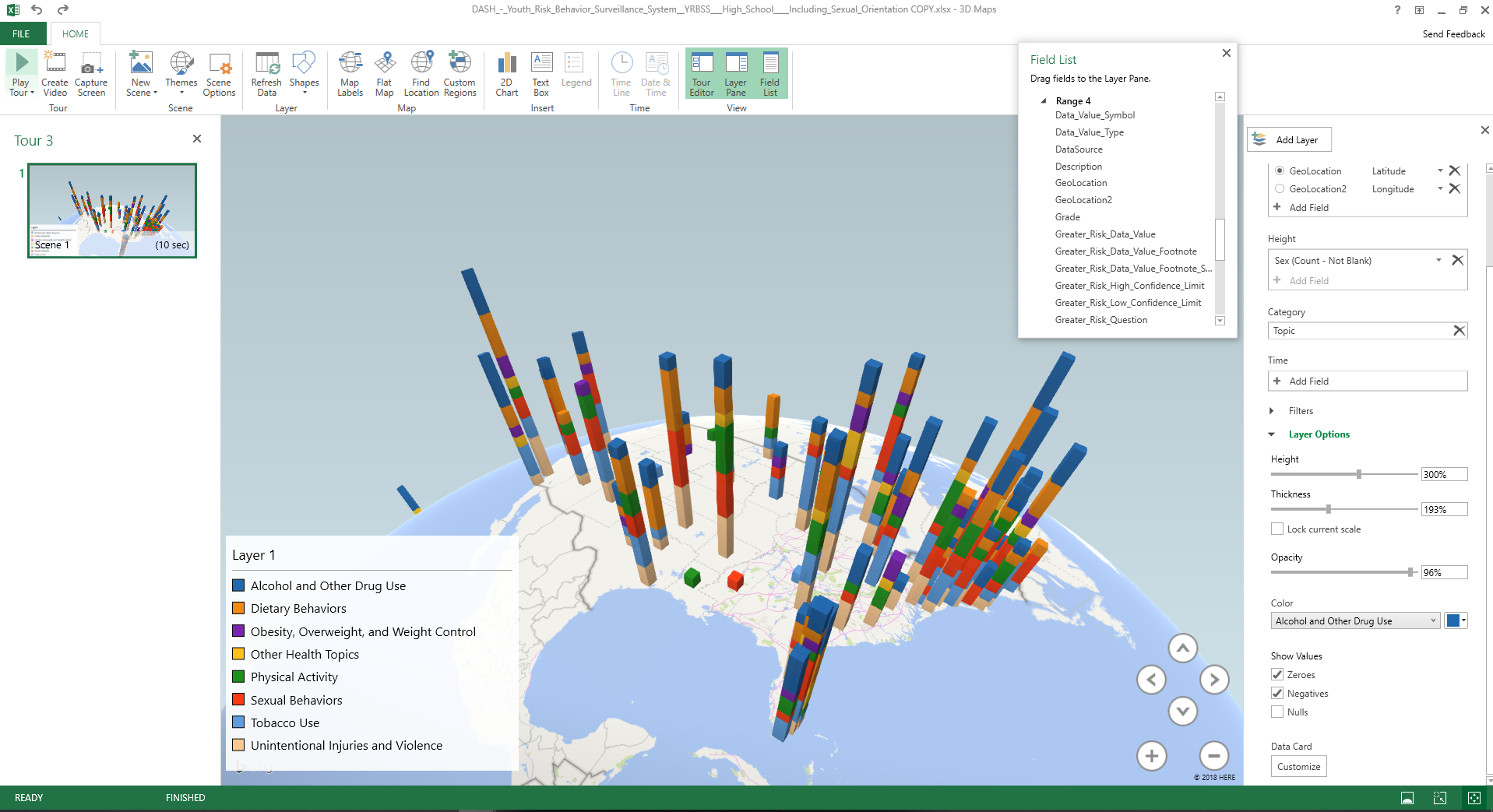
**Step 3: Visualization in Excel Sheet:**

*{Excel file can get corrupted and simple fix by open excel go to*

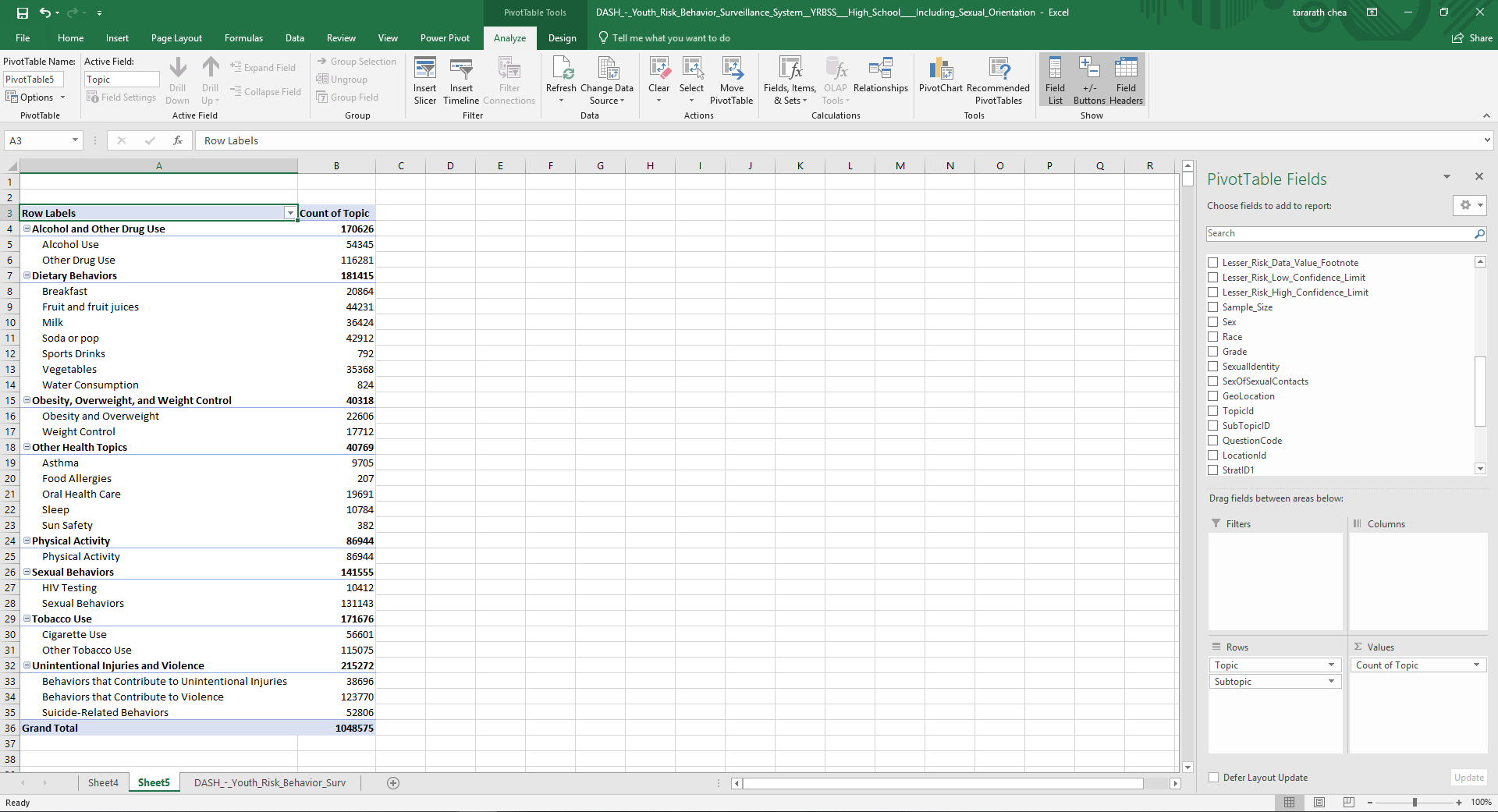
*file->options->add-ins->managed-> Go to dropdown box and check disabled items -> enable file youth risk files and try to reload the dataset again.}*

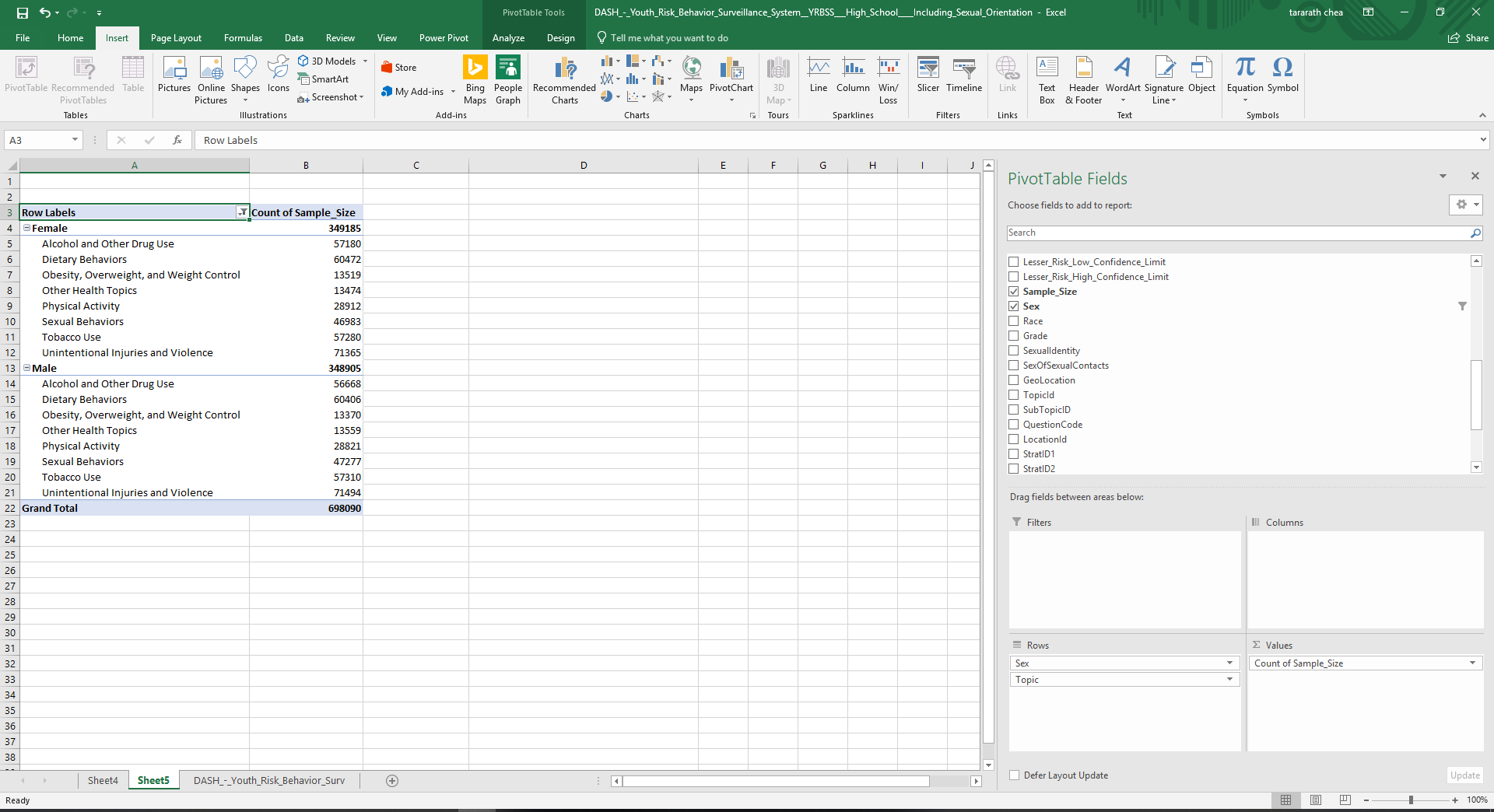
Visual graph from Youth Risk data in Excel

This 3D map is created from Excel sheet data that precisely focus on analyzing data to help improve the community, by showing the statistics of youth risk at health abuse. This visual graph show geolocation that that highly been reported in the data table. However, this visual may not exactly the same as the actual intended visual due to excel limitation on data retrieval application.

****

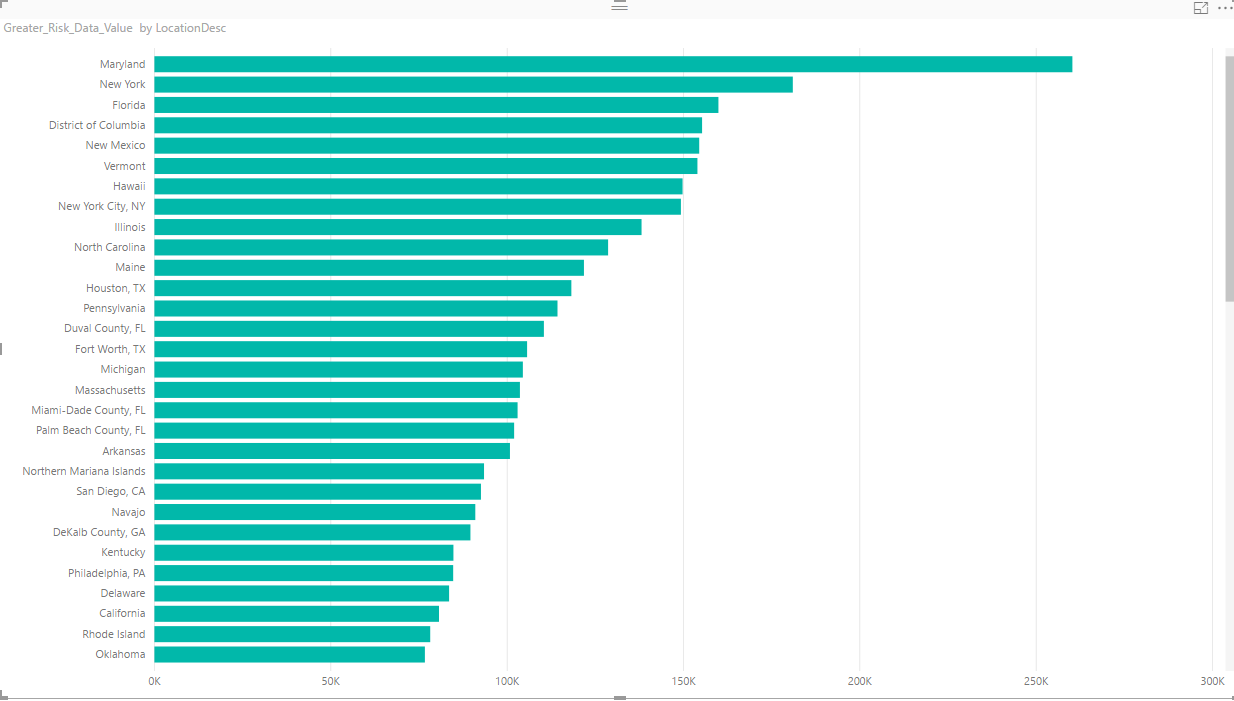
This is PivotTable created by the Youth Risk data fields. This table is a sample of data that we choose specifically from a total number of topic according to each risk. The second data table is separated by the total count based on their genders. This PivotTable can easily choose a specific category according to the user.



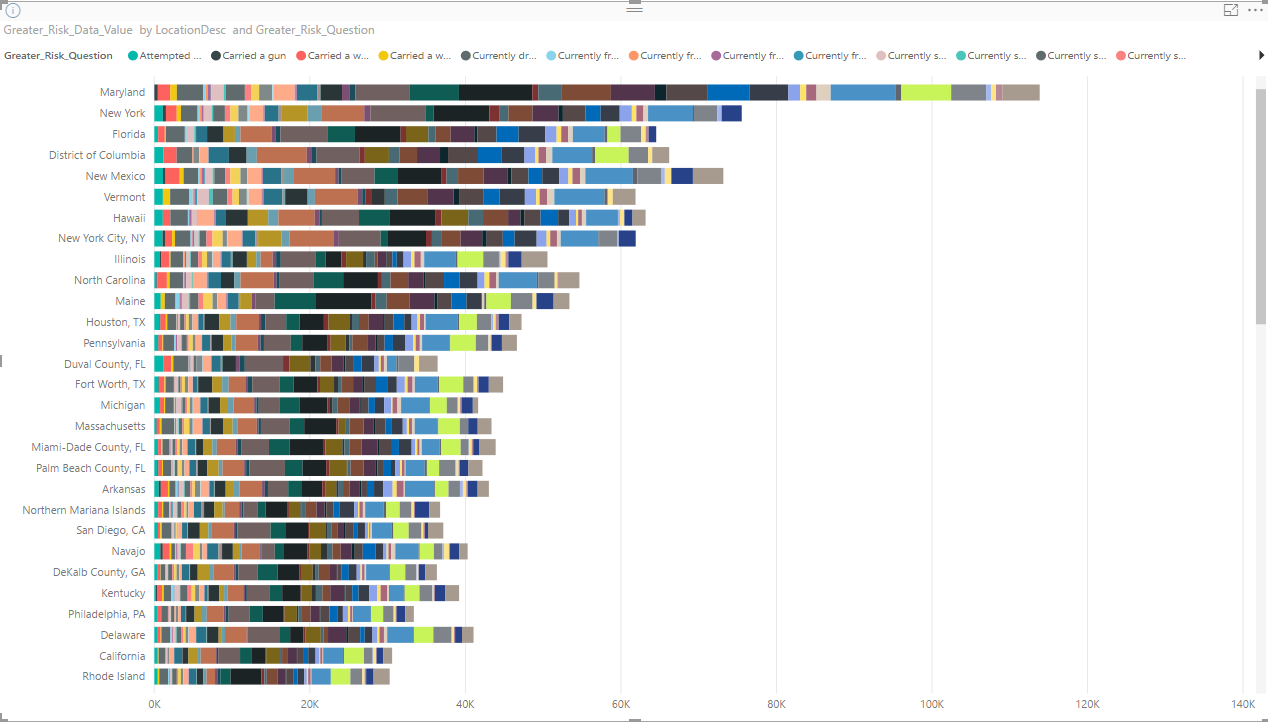


**Power Bi graphs data**

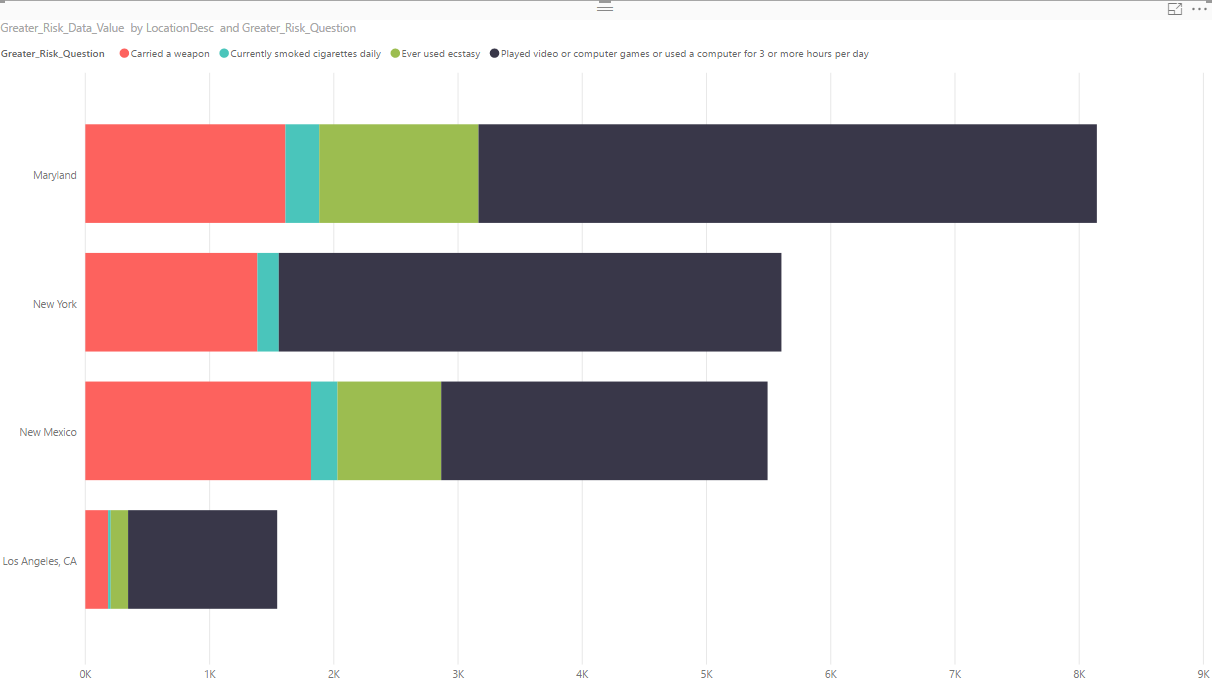
Why do we use Power BI? Power BI allows to view and analyze visual bigger data that cannot be opened in Excel. Power Bi used a powerful compression algorithm to import and cache the data. A large dataset can be easily cut down in size and aggregated to show more analysis.

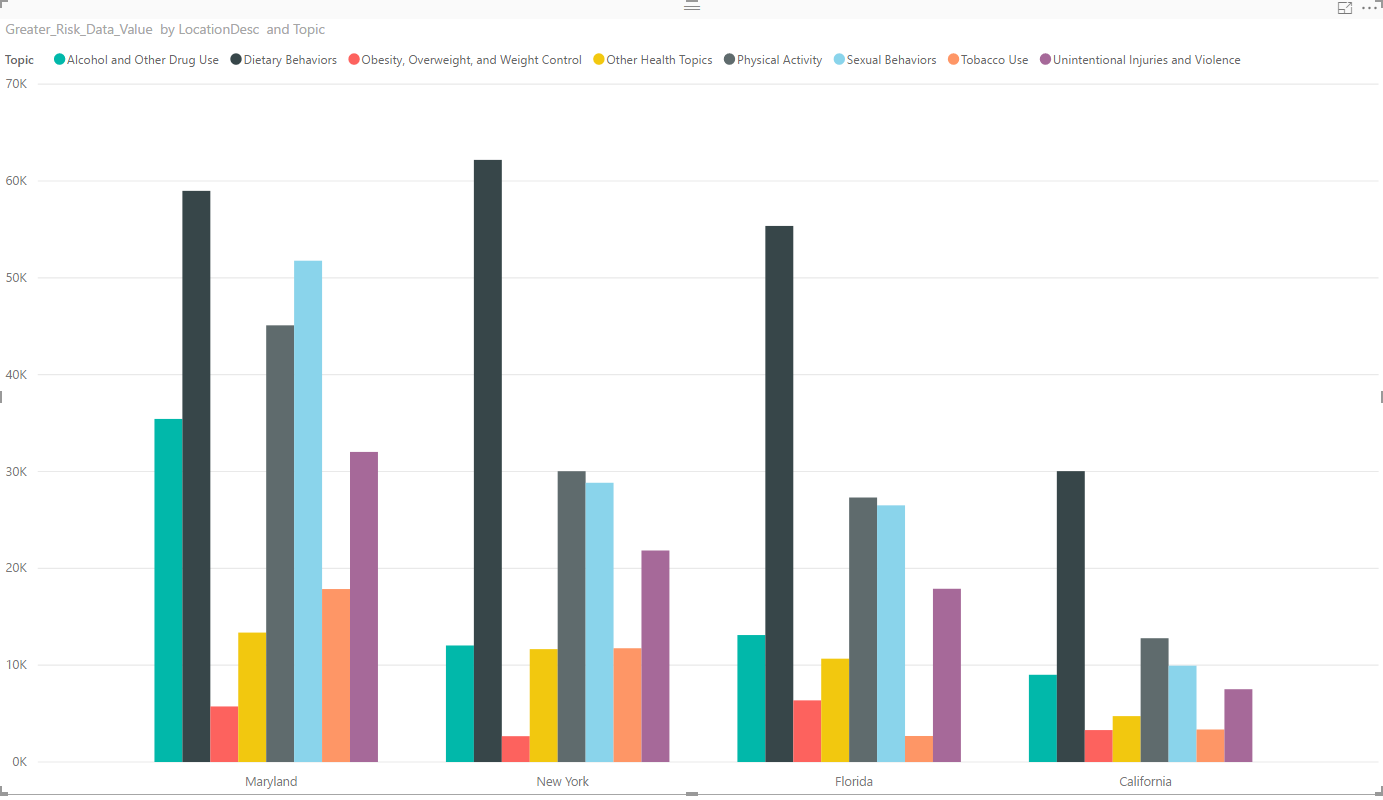
Graph data from greater risk data value table and Location destination 

Risk data graphs based on a topic covered



Risk data graph comparison of City Los Angeles to the top three states that have the highest risk data in the United States.

.

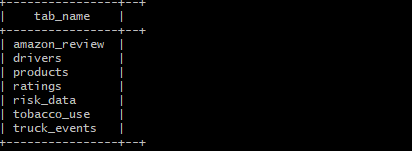


**PROBLEM ENCOUNTER**

* Data field corrupted from the download sources
* CSV file wasn’t completely open with Excel Sheet
* Excel can only open up 1,048,576 rows by 16,384 columns
* The data file was too big which resulted in missing data
* When creating the table in Beeline, table name has to be exactly the same as the given file
* Problem with creating a specific table to extract only tabacco\_use data. Below are the plans that were going to be applied to the table data to create tobacco\_use data. We were able to create the table, but we weren’t able to extract the field into this separate table.

*To focus on the topic “Tobacco\_Use”, we will first need to create a table that is filtered to only contain that Tobacco\_Use.*

*0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> CREATE TABLE IF NOT EXISTS Tobacco\_Use ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' AS select \* from risk\_data where topic = 'Tobacco Use';*



*0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> DROP TABLE IF EXISTS risk\_table;*

*--create the severity table by selecting from the risk\_table table*

*CREATE TABLE risk\_table*

*ROW FORMAT DELIMITED FIELDS TERMINATED BY ','*

*STORED AS TEXTFILE LOCATION '/user/pnguye47/youthrisk/tobacco\_use'*

**REFERENCES**

* Data URL: <https://catalog.data.gov/dataset/77ef7f4a-f208-4e52-bec2-53d349cb2375/resource/1b585349-4966-487c-ae3d-c4857c110cba>
* GitHub URL: <https://github.com/tchea/CIS4560-YouthRisk>
* Tutorial on Power BI: <https://www.techrepublic.com/blog/microsoft-office/how-to-download-and-install-microsoft-power-bi-desktop/>
* Download Power BI: <https://powerbi.microsoft.com/en-us/desktop/>