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

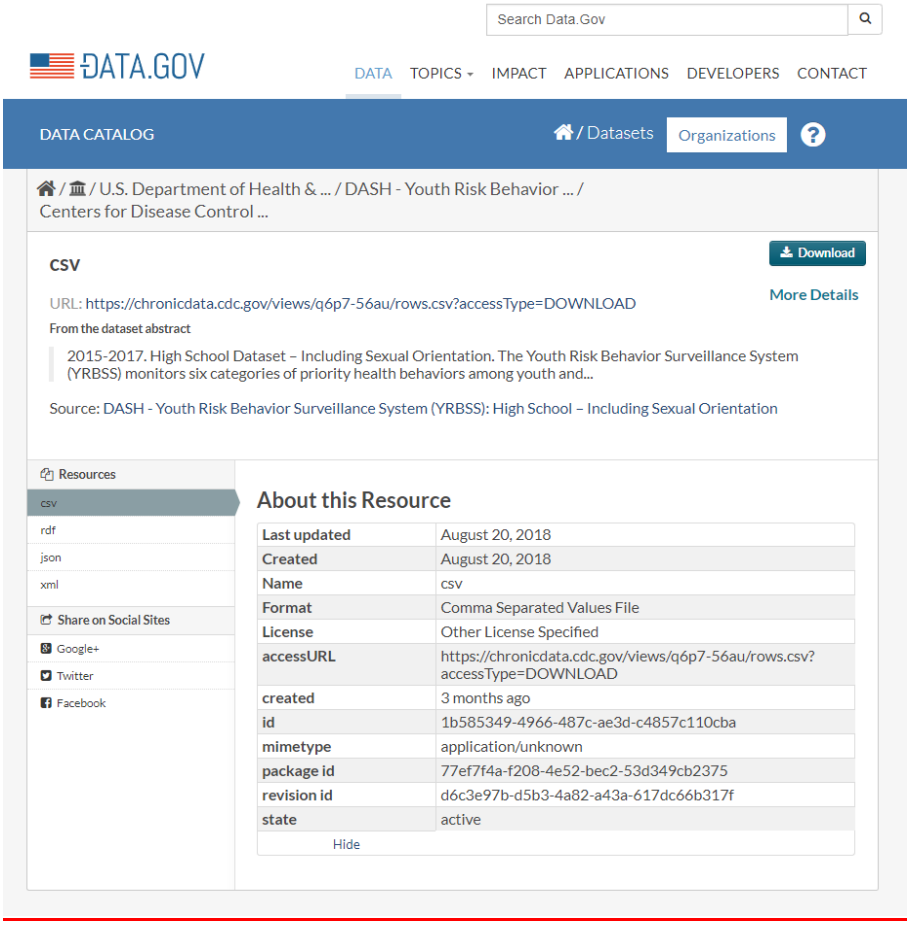
Objectives

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

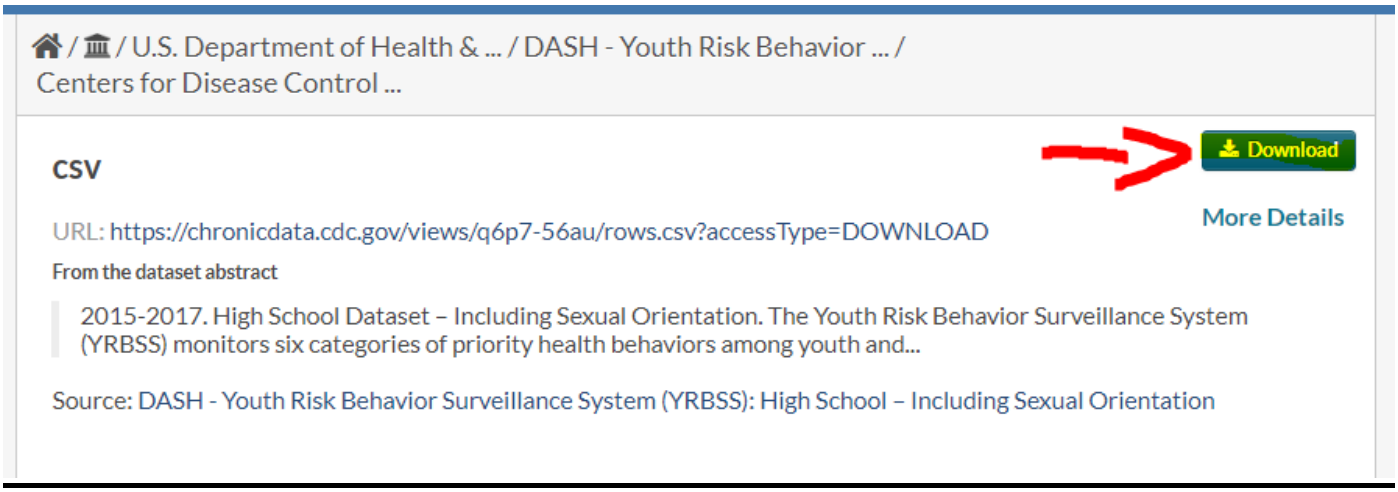
- How to download data from Data.gov
- Input data into Hadoop cluster
- Create a data table in Beeline
- Analyzing table with SQL commands
- Visualization
- Problem Encounters
- GITHUB Tutorial URL: https://github.com/tchea/CIS4560-YouthRisk

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

```
-bash-4.1$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                4
On-line CPU(s) list:   0-3
Thread(s) per core:    1
Core(s) per socket:    4
Socket(s):             1
NUMA node(s):          1
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 79
Model name:            Intel(R) Xeon(R) CPU E5-2699C v4 @ 2.20GHz
Stepping:              1
CPU MHz:               2195.294
BogoMIPS:              4390.35
Hypervisor vendor:     Xen
Virtualization type:   full
L1d cache:             32K
L1i cache:             32K
L2 cache:              256K
L3 cache:              56320K
NUMA node0 CPU(s):     0-3
-bash-4.1$
```

```
-bash-4.1$ hdfs dfs -df -h
Filesystem            Size     Used  Available  Use%
hdfs://mycluster  196.3 G  104.7 G     59.8 G   53%
-bash-4.1$
```

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

2. Try the following HDFS command to see if the file is in the server:
   -bash-4.1$ hdfs dfs -ls

3. Download the csv file onto the server:
   $ wget -O youthrisk https://chronicdata.cdc.gov/views/q6p7-56au/rows.csv?accessType=DOWNLOAD

```
perox@DESKTOP-LAUB4E1 MINGW64 ~
$ ssh pnguye47@129.150.205.28
pnguye47@129.150.205.28's password:
-bash-4.1$ wget -O youthrisk
wget: missing URL
Usage: wget [OPTION]... [URL]...

Try 'wget --help' for more options.
-bash-4.1$ wget -O youthrisk https://chronicdata.cdc.gov/views/q6p7-56au/rows.csv?accessType=DOWNLOAD
--2018-12-02 06:03:31--  https://chronicdata.cdc.gov/views/q6p7-56au/rows.csv?accessType=DOWNLOAD
Resolving chronicdata.cdc.gov... 52.206.140.205, 52.206.68.26, 52.206.140.199
Connecting to chronicdata.cdc.gov|52.206.140.205|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/csv]
Saving to: "youthrisk"

    [                       <=>                                                     ] 1,198,865,785 5.20M/s   in 3m 40s
```

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

   -bash-4.1$ hdfs dfs -mkdir youthrisk

   -bash-4.1$ hdfs dfs -put youthrisk youthriskdata

5. 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;

topic	total
YRBSS	1116428
Unintentional Injuries and Violence	322140
Dietary Behaviors	271335
Alcohol and Other Drug Use	256894
Tobacco Use	251138
Sexual Behaviors	210842
Physical Activity	131469
"Obesity	61460
Other Health Topics	56986

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;

topic
"Obesity
Alcohol and Other Drug Use
Tobacco Use
Dietary Behaviors
YRBSS
Sexual Behaviors
Physical Activity
Unintentional Injuries and Violence
Other Health Topics

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;

subtopic	total
Food Allergies	506
Sun Safety	1047
Water Consumption	2026
Sports Drinks	2093
Asthma	13297
HIV Testing	14555
Sleep	15059
Oral Health Care	27077
Breakfast	30183
"Obesity	41972

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;

locationdesc	topic	_c2
California	Other Health Topics	2053
California	Tobacco Use	6841
California	Physical Activity	4024
California	Sexual Behaviors	6643
California	Unintentional Injuries and Violence	10195
California	Dietary Behaviors	9145
California	Alcohol and Other Drug Use	8549
California	"Obesity	2075

8 rows selected (14.314 seconds)

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;

subtopic	total
Food Allergies	506
Sun Safety	1047
Water Consumption	2026
Sports Drinks	2093
Asthma	13297
HIV Testing	14555
Sleep	15059
Oral Health Care	27077
Breakfast	30183
"Obesity	41972
Other Health Topics	47371
Milk	48907
Vegetables	57469
Behaviors that Contribute to Unintentional Injuries	59842
Overweight	61460
Soda or pop	64914
Fruit and fruit juices	65743
Alcohol Use	79344
Suicide-Related Behaviors	79923
Cigarette Use	83264
Other Tobacco Use	167874
Alcohol and Other Drug Use	177204
Other Drug Use	177550
Behaviors that Contribute to Violence	182375
Tobacco Use	187650
Dietary Behaviors	192604
Physical Activity	222127
Unintentional Injuries and Violence	228498
Sexual Behaviors	346758

29 rows selected (41.408 seconds)

Short Question from the survey

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

```
+-------------------------------------------------------+
|                   shortquestiontext                   |
+-------------------------------------------------------+
| Daily breakfast eating                                |
| Ever cigarette use                                    |
| Frequent cigarette use                                |
| Fruit consumption >= 3 times                          |
| Milk drinking >= 2 glasses                            |
| Muscle strengthening                                  |
| Soda drinking >= 2 times                              |
| Soda or pop                                           |
| Television watching                                   |
| Vegetables                                            |
|  Overweight                                           |
| "Current cigarette                                    |
| Concussion                                            |
| Ever marijuana use                                    |
| Food allergies                                        |
| Initiation of cigarette smoking                       |
| Milk                                                  |
| No fruit consumption                                  |
| Sad or hopeless                                       |
| Water drinking >= 2 glasses                           |
| Current alcohol use                                   |
| Current marijuana use                                 |
| Current smokeless tobacco use                         |
| Current tobacco use                                   |
| Ever cocaine use                                      |
| Source of alcohol                                     |
| Sports team participation                             |
| Suicide plan                                          |
|  and Weight Control"                                  |
| Current daily smokeless tobacco use                   |
| Ever inhalant use                                     |
| Ever synthetic marijuana use                          |
| Fruit consumption >= 1 time                           |
| Fruit consumption >= 2 times                          |
| HIV Testing                                           |
| No soda drinking                                      |
| No vegetable eating                                   |
| Physical fighting at school                           |
| Weapon carrying at school                             |
| "Shot                                                 |
| Behaviors that Contribute to Violence                 |
| Breakfast                                             |
| Condom use                                            |
| Current daily electronic vapor product use            |
| Drive when using marijuana                            |
| Initiation of marijuana use                           |
| Physical activity                                     |
| "Pill                                                 |
| Behaviors that Contribute to Unintentional Injuries   |
| Current cigarette use                                 |
| Daily PE attendance                                   |
```

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;

genderethnicity	total
	1008724
Total	251300
Female	194342
Male	194160
0	80554
Native Hawaiian or Other Pacific Islander	57486
White	57421
Asian	57338
Hispanic or Latino	57275
Black or African American	57241

10 rows selected (16.953 seconds)

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;

locationdesc	total
United States	59624
West Virginia	53111
"Fort Worth	53042
"Houston	52914
"Orange County	52871
"Broward County	52708
Oklahoma	52703
"Miami-Dade County	52673
"Palm Beach County	52662
Northern Mariana Islands	52603

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;

topic	subtopic	total
YRBSS	Unintentional Injuries and Violence	228498
Sexual Behaviors	Sexual Behaviors	196287
YRBSS	Dietary Behaviors	192604
YRBSS	Tobacco Use	187650
Unintentional Injuries and Violence	Behaviors that Contribute to Violence	182375
Alcohol and Other Drug Use	Other Drug Use	177550
YRBSS	Alcohol and Other Drug Use	177204
Tobacco Use	Other Tobacco Use	167874
YRBSS	Sexual Behaviors	150471
Physical Activity	Physical Activity	131469
YRBSS	Physical Activity	90658
Tobacco Use	Cigarette Use	83264
Unintentional Injuries and Violence	Suicide-Related Behaviors	79923
Alcohol and Other Drug Use	Alcohol Use	79344
Dietary Behaviors	Fruit and fruit juices	65743
Dietary Behaviors	Soda or pop	64914
"Obesity	Overweight	61460
Unintentional Injuries and Violence	Behaviors that Contribute to Unintentional Injuries	59842
Dietary Behaviors	Vegetables	57469
Dietary Behaviors	Milk	48907

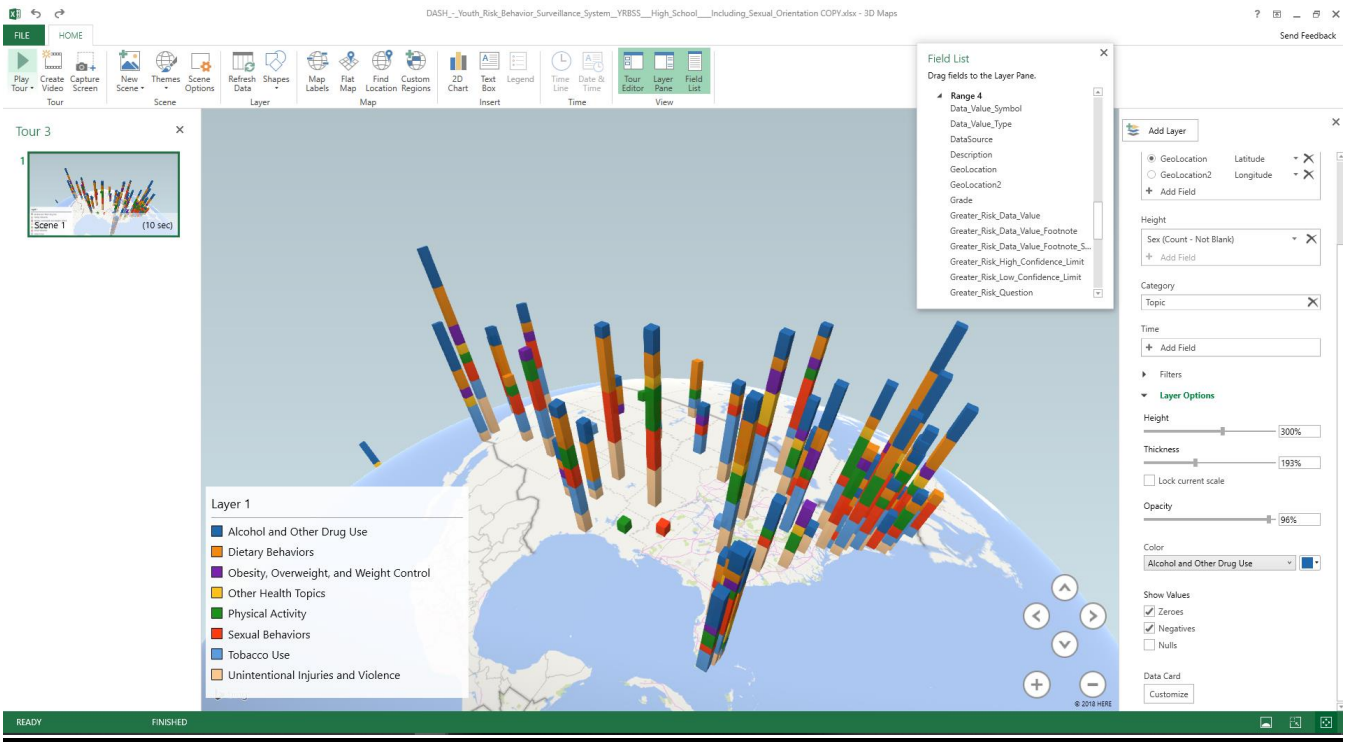
20 rows selected (23.983 seconds)
0: jdbc:hive2://cis4560-bdcsce-4.compute-6082>

Step 3: Visualization in Excel Sheet:

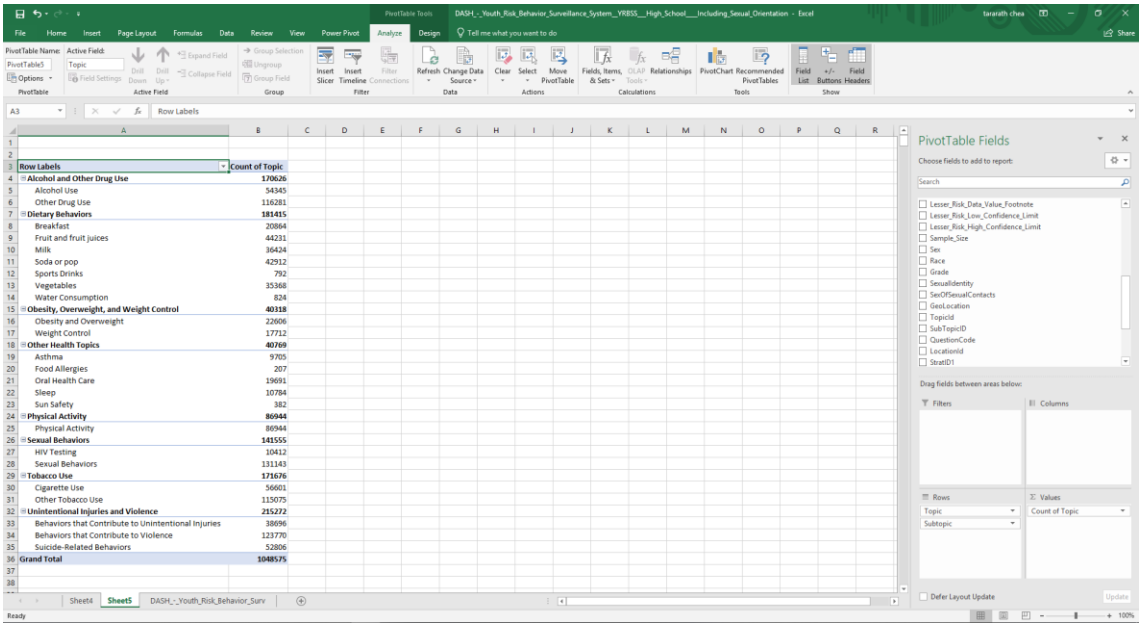
{Excel file can get corrupted and simple fix by open excel go to file->options->add-ins->managed-> Go to dropdown box and check disabled items -> enable file youth risk files and try to reload the dataset again.}

Visual graph from Youth Risk data in Excel

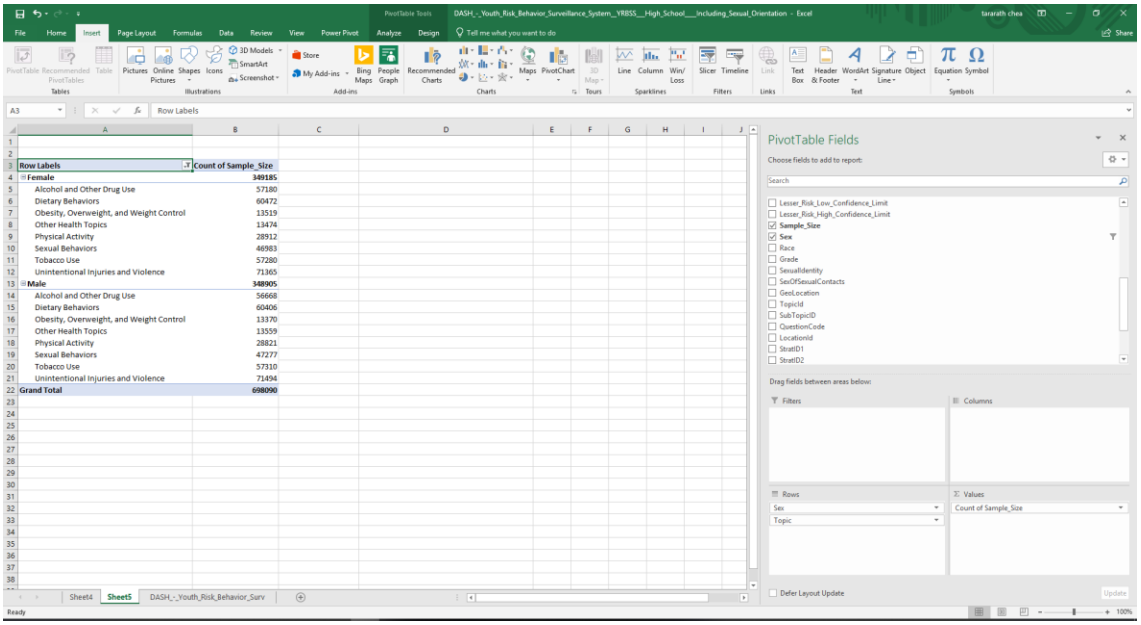
This 3D map is created from Excel sheet data that precisely focus on analyzing data to help improve the community, by showing the statistics of youth risk at health abuse. This visual graph show geolocation that that highly been reported in the data table. However, this visual may not exactly the same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



Row Labels	Count of Topic
Alcohol and Other Drug Use	170626
Alcohol Use	54345
Other Drug Use	116281
Dietary Behaviors	181415
Breakfast	20864
Fruit and fruit juices	44231
Milk	36424
Soda or pop	42912
Sports Drinks	792
Vegetables	35368
Water Consumption	824
Obesity, Overweight, and Weight Control	40318
Obesity and Overweight	22606
Weight Control	17712
Other Health Topics	40769
Asthma	9705
Food Allergies	207
Oral Health Care	19691
Sleep	10784
Sun Safety	382
Physical Activity	86944
Physical Activity	86944
Sexual Behaviors	141555
HIV Testing	10412
Sexual Behaviors	131143
Tobacco Use	171676
Cigarette Use	56601
Other Tobacco Use	115075
Unintentional Injuries and Violence	215272
Behaviors that Contribute to Unintentional Injuries	38696
Behaviors that Contribute to Violence	123770
Suicide-Related Behaviors	52806
Grand Total	1048575

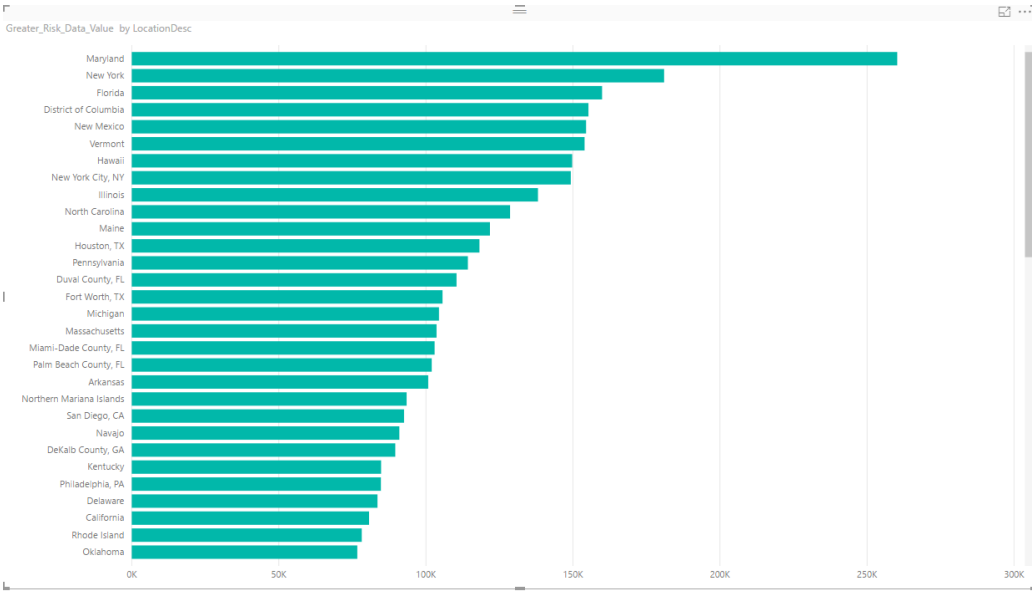


Row Labels	Count of Sample_Size
Female	349185
Alcohol and Other Drug Use	57180
Dietary Behaviors	60472
Obesity, Overweight, and Weight Control	13519
Other Health Topics	13474
Physical Activity	28912
Sexual Behaviors	46983
Tobacco Use	57280
Unintentional Injuries and Violence	71365
Male	348905
Alcohol and Other Drug Use	56668
Dietary Behaviors	60406
Obesity, Overweight, and Weight Control	13370
Other Health Topics	13559
Physical Activity	28821
Sexual Behaviors	47277
Tobacco Use	57310
Unintentional Injuries and Violence	71494
Grand Total	698090

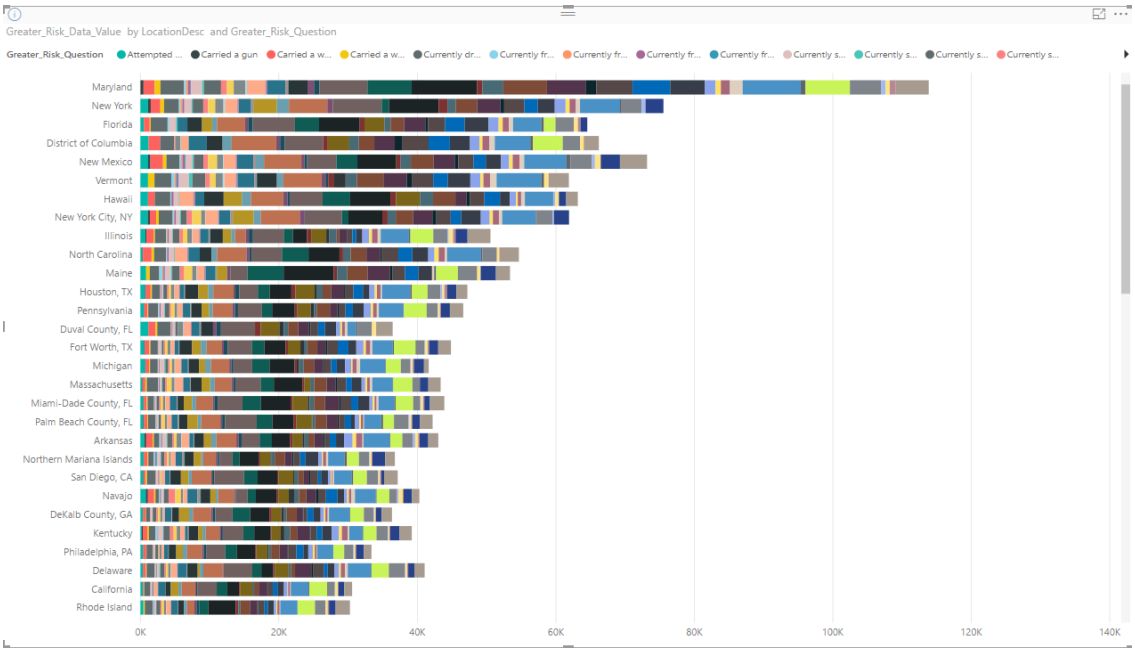
Power Bi graphs data

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

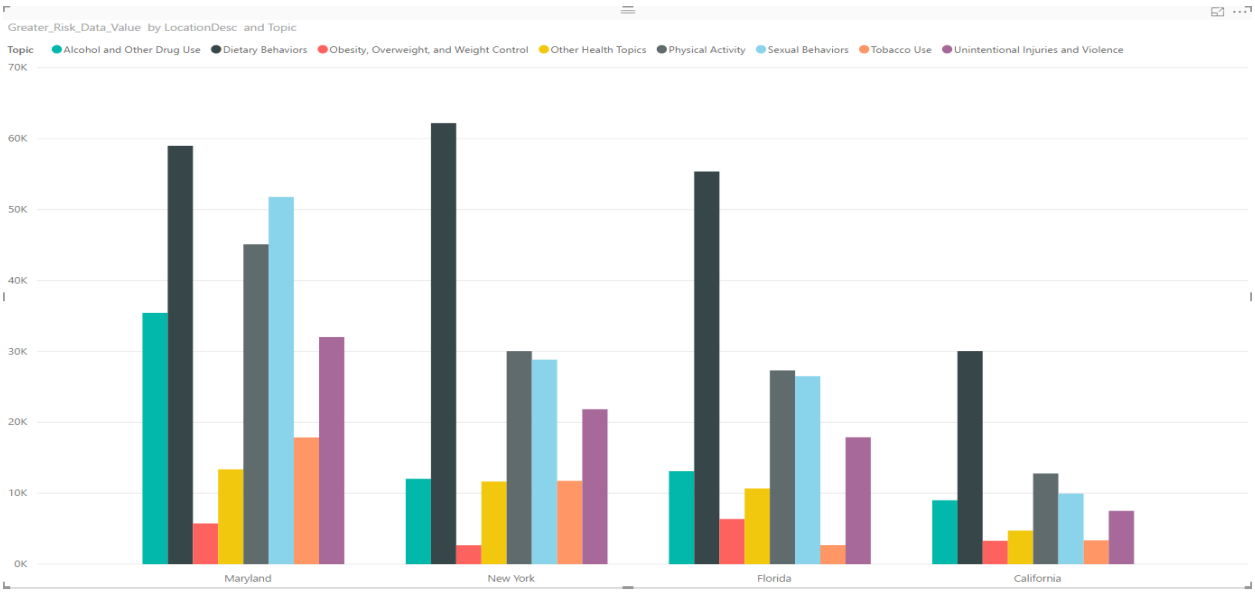


Risk data graphs based on a topic covered



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';*

```
+-----------------+--+
|    tab_name     |
+-----------------+--+
| amazon_review   |
| drivers         |
| products        |
| ratings         |
| risk_data       |
| tobacco_use     |
| truck_events    |
+-----------------+--+
```

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/