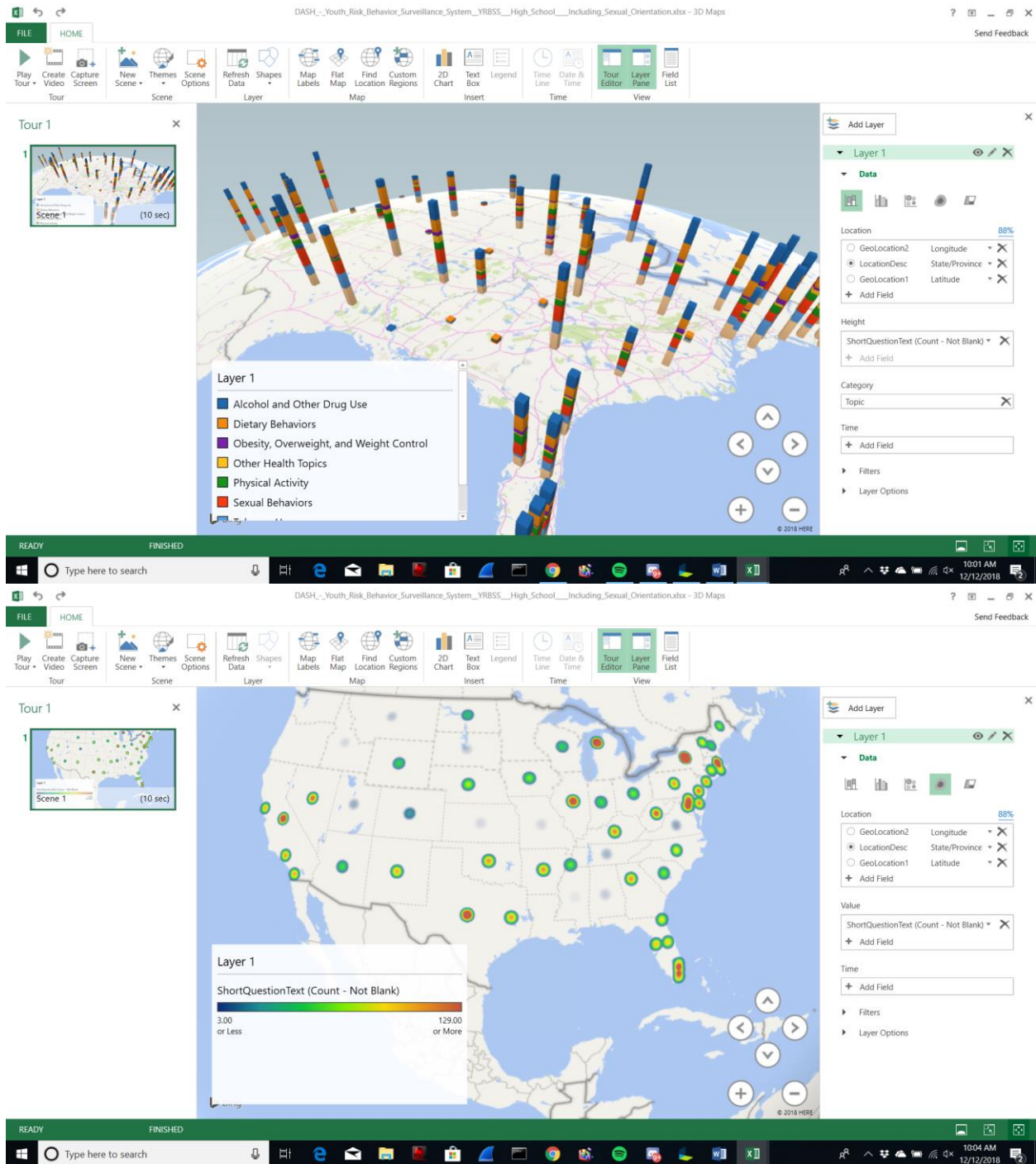


Youth Risk Statistics and Prediction



Abstract: The purpose of our research was to compile Youth Risk Behavior Statistics in order to do a comparison between all the different states and risk topics in order to predict which of these have the highest and lowest risk levels and based from the data. The United States was chosen in order to narrow down the information while at the same time keeping the information really relevant to the authors and people living in the United States. The data discovered, revealed many states and risk topics some that would be extremely useful for people to understand which states have a high risk for teenagers. We found our data from chronic data website which specializes in having datasets related to public health. These datasets are great for people who would like to know more about health throughout the world.

1. Introduction

Our research led us to many risk behavior statistics from 2015 and 2017 only. The data were recorded from chronic data website of the only two years of 2015 and 2017. We did an in-depth research of the data and found that this data is recorded through the school using surveys. The surveys can only be taken down if parents or guardians agree to it. While working on the dataset we realize that it was the ideal size and found that the information available would prove to be valuable to most of the people who are trying to figure out the behavior risk levels throughout the states in the United States. This data may be useful for people who are curious about this topic. It basically gives the viewers a better insight of what teenagers' experience in school and these surveys are only taken down from guardians/parents who approve the survey. These data may reflect on students who are feeling depressed or suicidal. This is a great system to rely on and it can save many lives when the school decides to take action fast. Mass shooting and suicidal have always been a big problem that has affected many people's lives. Schools can improvise a system that helps students' lives, encouraging them to know that life is important. If they ever feel lonely or have family problems they can always talk to their school counselor. This data This information can help prevent students from doing risky behaviors if we

give them a survey. To help students in need of special therapy treatment. With this knowledge, we can dig deeper on why students may feel like doing risky behaviors. The top risky behavior that teenagers commit regarding from the visual data is alcohol and drug usage.

2. Analysis Model

Our approach to analyzing the data can be broken down into six major steps: Data Gathering, Extraction of compressed data, compiling of data, table creation through HiveQL, and our analysis through the information in front of us. Each step will be touched on below and is visualized in figure 1 below.

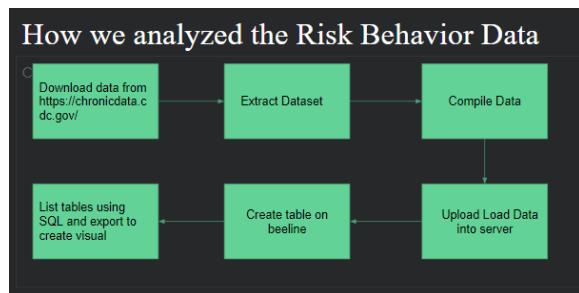
The first step I mentioned was data gathering, and this step was an extremely difficult step. This is so because of the sheer amount of data available across the vast and many platforms. We knew from the start what data we wanted to create an analysis for and that was through our interest in Youth Behavior Risk Surveillance System (YRBSS). Searching through different data repositories and warehouses was a difficult task, going through sites such as google aws, kaggle, socarata, and then finally settling with CRC.gov CRC.gov linked us to the dataset on YRBSS. After downloading the data, we had to decompress the file.

The second step we took was the extraction of compressed data. The files we downloaded were compressed because of the sheer size of the information, and in order to do so, we had to extract the data in order to narrow down the topics to giving us the opportunity to compile them in topics in order to do comparative analysis. We worked with the years of 2015 and 2017 to keep the data relevant in the present day because that was the only data that was given to use for the data. To compile the data, we chose to group the data by topic and states that have the highest to lowest results. We had to search through each dataset to confirm each of them had the information we were looking for. As soon as it was confirmed, we proceeded to clean the data by removing any null columns, and any properly labeling the important ones. We just made sure all the important aspects of data cleaning were checked: Topics, total, completeness, locationdesc, and

uniformity. This was an extremely important step so; the data was much easier to navigate when uploaded. After cleaning and compiling the data, we then proceeded to upload the data to Gitbash server that was assigned by Professor Woo.

Uploading the data to Gitbash server was the much more straightforward step. We ran the typical queries `-wget -O youthrisk` (to download the data on the big data platform), after extracted the file from a folder `-mkdir` (to create a directory for the data in order to make the data much more easily accessible, and finally we ran to `-ls` commands to confirm that the data was in the correct directory. After confirming all the data was uploaded we proceeded to run Hive in order to create a table to structure the information.

In order to create the table, we had to choose all the relevant columns in each table, so we could add structure to all of the information. By structuring the data, we made it much more digestible to each of the group members who wanted to run queries. We structured the data by making each column definitions specific in order to narrow our research target, specifying the file formats, and data types. After creating the tables, we then proceeded with our analysis, which will be touched on.



(Figure 1: Our analysis depicted.)

3. Data Analysis

Youth Risk Behavior data usually consist of a collection of risk topics taken down by a survey given by the school. Each risk topic correlates to reasons why students may be drinking alcohol or doing drugs. From looking at the data peer pressure is a problem in high school so they may feel obligated to do it. This data set is comprised of uniquely risk

topics in the following year of 2015 and 2017 only, each also containing latitude, longitude of where the states are being recorded down for risk behavior towards teenagers. In this paper, we are interested in analyzing Youth Risk Behavior Surveillance System. Based on our data we can find Oklahoma, Rhode Island, and California had the that the lowest risk level. The highest risk level was in Maryland. When comparing California to Maryland it made California look so small with behavior risks. The data was not able to have a timeline lack of dates and time. It only had years of 2015 and 2017. We tried everything it was struggling, but this is the most we can do for the Geotimes.

Abstract: Type the abstract at the beginning of the first column. The abstract should be no longer than 200 words.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in this example.

2.2 Sections

Headings: Type and label section and subsection headings in the style shown on these pages. Use numbered sections, in order to facilitate cross references.

References: Citations within the text appear in brackets as [ref. number]. Gather the full set of references together under the heading References; place the section before any Appendices, unless they contain references. Arrange the references in the order that they are cited in the text. Provide as complete a citation as possible, using a consistent format.

Appendixes: Appendixes, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: Appendix A Title of Appendix.

2.3 Footnotes

Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.^[1] Footnotes should be separated from the text by a line.^[2]

2.4 Graphics

Illustrations: States, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions for figures below the figures. Type the captions for tables above the tables.

3. Length of Camera-ready Manuscript

For the length of camera-ready manuscripts, **a paper is limited up to 4 pages**. All illustrations, references, and appendices must be accommodated within these page limits. Any extra page beyond the first four pages will be deleted. Please DO NOT put a page number in each page

4. Submission Process

1. Format your paper using this template.
2. Turn the hardcopy by Dec 4th before the lecture starts

References

- [1] T.A. Jones, "Writing a good paper," *IEEE Trans. on General Writing*, Vol. 1, no. 2, pp.1-10, May 2002.
- [2] K. Hwang, *Computer Arithmetic*, John Wiley, 1997.

¹ This is how a footnote should appear

² Note the line separating the footnotes from the text