Data Story

Anitha Pissay

October 23, 2018

**Introduction**

The **Youth Risk Behavior Surveillance System (YRBSS)** was developed in 1990 to monitor health behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. These behaviors, often established during childhood and early adolescence, include

* Behaviors that contribute to unintentional injuries and violence.
* Sexual behaviors related to unintended pregnancy and sexually transmitted infections, including HIV infection.
* Unhealthy dietary behaviors.
* Inadequate physical activity.

My area of focus will be based on

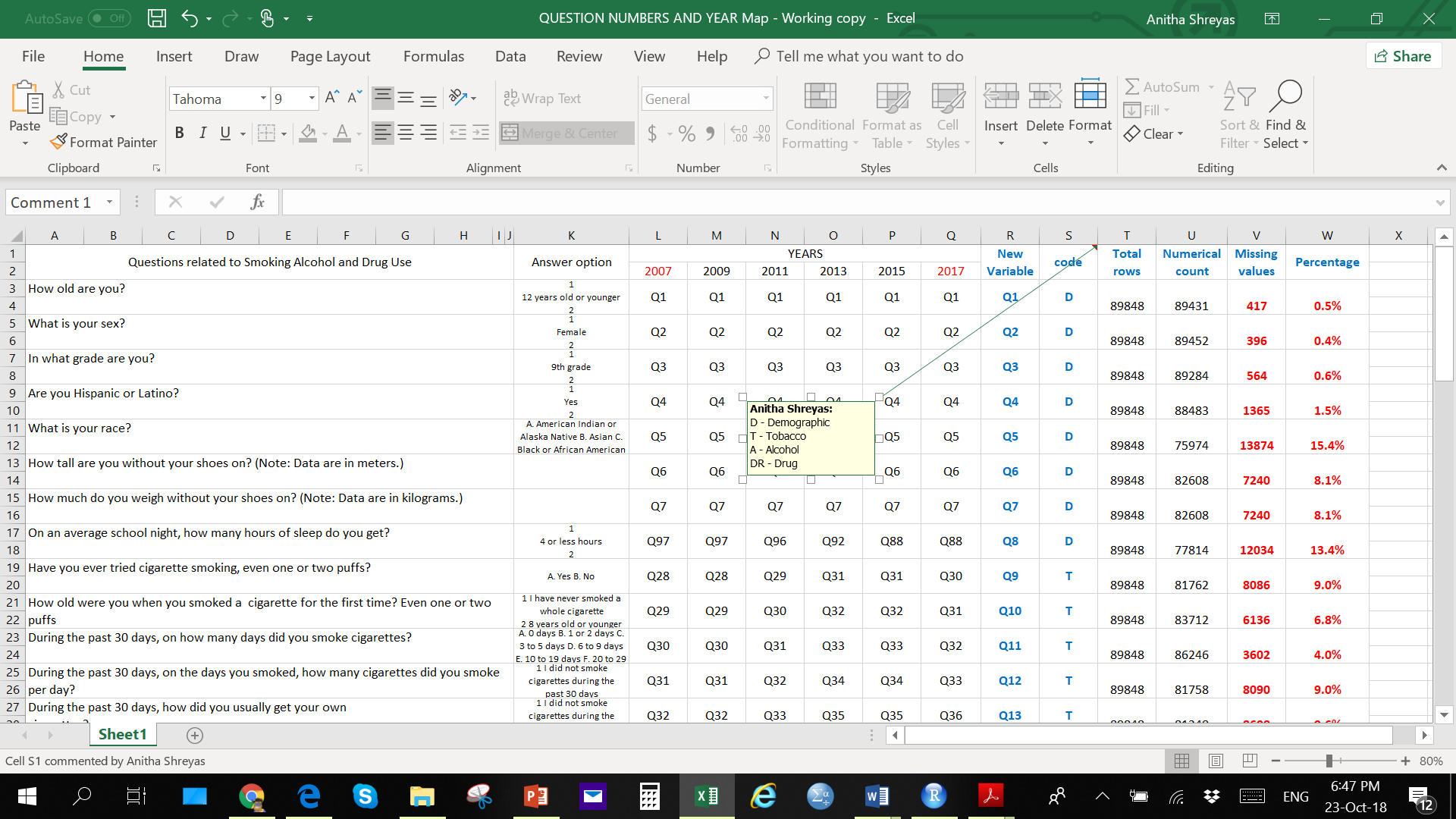
* **Tobacco use.**
* **Alcohol and other drug use.**

**Data Collection**

From 1991 through 2017, the YRBSS has collected data from more than 4.4 million high school students in more than 1,900 separate surveys. YRBSS data sources include ongoing surveys as well as one-time national surveys, special-population surveys, and methods studies. The ongoing surveys include school-based national, state, tribal, and large urban school district surveys of representative samples of high school students and, in certain sites, representative state, territorial, and large urban school district surveys of middle school students. My project primarily focuses on **Tobacco, Alcohol and other drug Use** **(TAD)** from 2007 to 2017. The Data sets were downloaded from [“Center for Disease Control and Preventions”](https://www.cdc.gov/healthyyouth/data/yrbs/data.htm) website.

**Data Wrangling**

1. My approach in solving this problem is by collating all the data collected from 2007 to 2017 from the individual years to create a single cumulative project data file with matching variables.
2. As the area of focus was narrowed to only 3 behaviors – TAD, only the responses of the questions related to them were extracted along with the demographic data such as students age, height, weight, etc.
3. The challenge was that questions asked in the survey were *numbered differently* in each data file across different years, therefore posing a greater risk in errors while merging. Hence the questions had to be visually mapped first amongst these files before merging them onto the new datafile. For this purpose, a new set of variables were created, and the responses added as per the [map](https://github.com/anitha148/Springboard_Projects/blob/master/Capstone%20Projects/Project_YRBSS/QUESTION%20NUMBERS%20AND%20YEAR%20Map.xlsx) to ensure the accuracy of the data.



A cropped screen short of the mapping file used to create the datafile

1. An extra variable (YEAR) that did not belong to any of the datafiles was created to identify the year from which the data was being extracted.
2. For ease in data readability and understanding appropriate labels were added for these values during the visualization methods and coding techniques were applied.
3. The variables are labelled **Year** and **Q1 to Q38** within which

* ‘**Year**’ identifies the year from which the responses in each row were extracted from.
* responses to questions related to demographics are in Q1 to Q8
* responses to questions related to Tobacco are in Q9 to Q19
* responses to questions related to Alcohol are in Q20 to Q25
* responses to questions related to drug use are in Q26 to Q38

Q14 and Q15 asks about electronic cigarettes also known as e-cigarette and has data only for year 2015 and 2017. The entire data was merged on an excel file which was later converted to a .csv file.

1. The final datafile contains approximately 88000 rows with missing values in certain rows and columns for different variables. The responses are recorded as categorical variables.
2. In R studio, the csv file was imported, and data wrangling method was applied starting with replacing the NA (missing values) with ‘0’ as all the responses (except for Demographic responses) were recorded as numeral values starting with 1.
3. For ease in visualization in outputs the values of the variables were converted to labels by applying data wrangling techniques.
4. Once data is wrangled into its desired observation format, it will then be concatenated to a cumulative data frame.
5. Further data wrangling will be used in order to calculate certain statistics, such as ratios, percentages, and trend, and to create data frames for constructing the predictive and classification models.

.