Big Data Analysis of Youth Tobacco Smoking Trends in the United States

Shilpa Balan[#1], Nishant Shristiraj[#2], Vrunda Shah[#3], Anusha Manjappa[#4]

[#]California State University-Los Angeles, College of Business and Economics, Department of Information Systems

[1]sbalan@calstatela.edu, [2]nshrist@calstatela.edu, [3]vshah16@calstatela.edu, [4]amanjap@calstatela.edu

Abstract— **As large amounts of data are now generated in the healthcare industry, big data technologies are used to process these data. Tobacco smoking has been significant among the youth in the United States. In this research, we use big data techniques such as R and Tableau to explore the tobacco smoking trends among the youth in the United States. Results indicate that there is more number of youth male smokers than youth female smokers. The results also indicate that more than 51 percent of current youth smokers want to quit smoking.**

Keywords-big data; tobacco; smoking; Tableau; R

I. INTRODUCTION

The healthcare industry has generated large amounts of data [18]. Data from the U.S. healthcare system reached 150 exabytes in 2011 [18]. With the rising volume of data, there is a need to apply big data technologies to healthcare data in order to analyze it. The analysis can help with the improvement of healthcare quality and increased patient satisfaction.

Big data in healthcare is now being used to predict epidemics, cure diseases, and avoid preventable deaths. With the increasing world population and the increasing life span, several decisions of patient treatment are driven by data [9]. The goal is to understand as much about a patient as possible [22].

Cigarette smoking is a leading risk factor for morbidity in the United States [4, 11, 12, 15]. Cigarette smoking has a significant effect on the health of Americans, and is a major cause of cardiovascular disease [8].

The Centers for Disease Control and Prevention (CDC) estimates that over 18.1% of adults are consistent smokers [1, 16]. Tobacco smoking is one of the leading causes of preventable death [10]. Smoking can harm the human body, and it can directly result in death from heart disease, cancers or strokes.

The main aim of this research is to analyze the tobacco smoking trends among the youth in the United States. This research also aims to increase the awareness of the health impacts of tobacco smoking.

II. RESEARCH BACKGROUND

Cigarette smoking is the leading cause of preventable disease and death in the United States that accounts for more than 480,000 deaths every year [23]. More than 16 million Americans live with a smoking-related disease [2].

Cigarette use has grown rapidly, increasing between the 1910's and the mid-1960's [24]. In the past, people in the United States consumed tobacco primarily in the form of chewing tobacco and cigars [24]. Although cigarette consumption has been declining, cigarettes have remained the most commonly used tobacco product in the United States [2]. Tobacco smoke is a mix of more than 7,000 chemicals [25]. The smoking rate has been falling for decades, but it usually drops only by a small percentage in a year [13].

Currently, cigarette smoking is the measure used to describe the prevalence of cigarette smoking [5]. Different definitions of current smoking have been defined. NHIS (National Health Interview Survey) defines current smoking as individuals having smoked at least 100 cigarettes during their lifetime and smoking every day or some days [14]. Current smoking for youth, young adults, and adults is defined as having smoked part or all of a cigarette during the past 30 days [14].

According to the (2012) NSDUH survey, the prevalence of current cigarette smoking among the youth 12–17 years of age was 6.6% and was similar among youth males (6.8%) and youth females (6.3%) [14]. From previous research it is seen that smoking prevalence is higher among adult males than among adult females, and that smoking is highest between ages 20 and 40 [7].

There is a need to improve knowledge of the dangers of smoking [18]. Cigarette smoking leads to diseases such as cardiovascular disease, chronic obstructive pulmonary disease and lung cancer [8]. Higher education is associated to higher awareness of the impacts of smoking [19].

In this research, the current tobacco smoking trends among the youth in the United States are analyzed using technology such as Tableau and R. These are further described in the Methodology section.

III. RESEARCH METHODOLOGY

The dataset analyzed in this research on tobacco smoking trends of the youth is an open dataset published by the Centers for Disease Control and Prevention [3]. The data contains information on smoking status, cessation, and cigarette use among the youth. Seventeen years of data from 1999-2016 was considered for the analysis in this research.

Figure 1 depicts the research methodology. The data was organized in a CSV format. Afterward, the data was then cleaned. The data was then analyzed using a technology named R.

R is an integrated suite of software facilities for data manipulation [17]. The data visualizations were further enhanced using a data visualization tool named Tableau. Tableau provides valuable insights with meaningful visualizations of the data. The visualizations generated with the help of Tableau are shown in Figures 2 and 3. This platform is easy to deploy and manage [21].

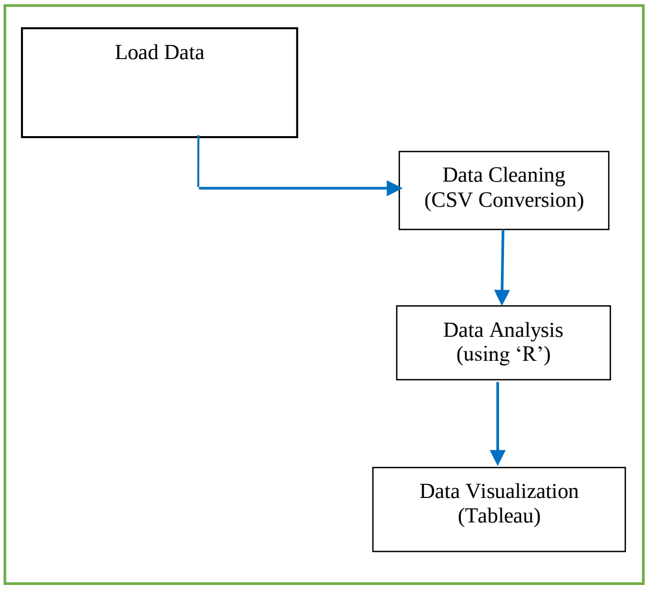

Load Data
Data Cleaning (CSV Conversion)
Data Analysis (using ‘R’)
Data Visualization (Tableau)

Figure 1. Architecture of Research Methodology

IV. RESULTS

This section describes the results of our study. Here, we present our most important results. The results provide valuable information on tobacco smoking trends among the youth. Figure 2 shows five states in United States that ranks highest for the number of youth tobacco smokers. West Virginia ranks first for the highest number of youth tobacco smokers. While there is not much variation in the number of tobacco smokers between Mississippi and Alabama, the state of North Carolina has a considerably lower number of youth tobacco smokers.

Figure 3 illustrates a comparison of youth tobacco smokers between male and female. There has been a constant decline in the number of youth male and youth female smokers.

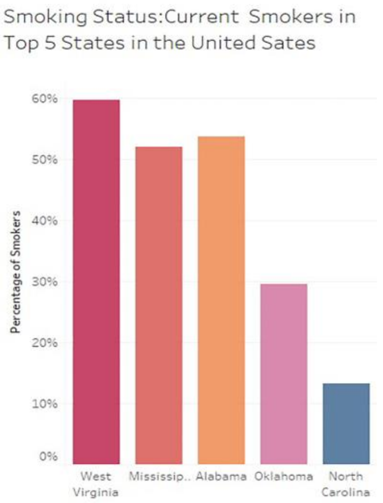


Figure 2. Highest Ranked States in United States for Youth Tobacco Smokers

From Figure 2, for the states in the United States that ranked highest for the number of youth tobacco smokers, we made a gender comparison of youth male and female tobacco smokers in Figure 3. It is found that the number of youth male smokers are slightly more than the number of youth female smokers. We were also able to extract summary statistics on the data using R. Figure 4 shows that more than 51% of the current youth smokers in the United States want to quit smoking.

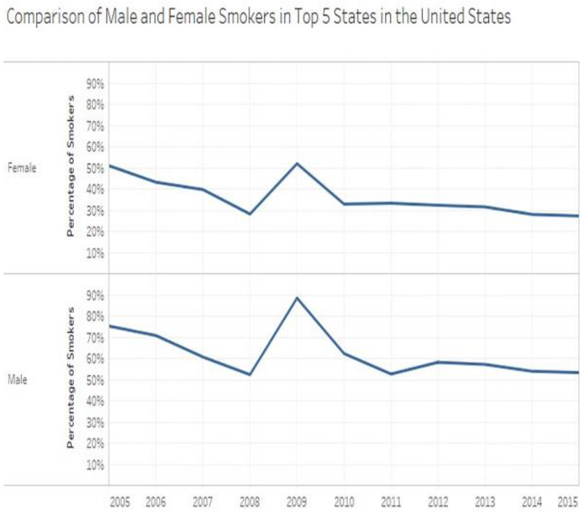
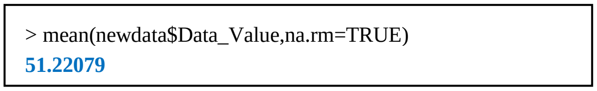


Figure 3. Male vs Female Comparison of Youth Tobacco Smokers



```
> mean(newdata$Data_Value,na.rm=TRUE)
51.22079
```

Figure 4. Summary Statistics for Percent of current Youth Smokers who want to Quit

Figure 5 shows the pseudocode of R program for computing the mean of the percent of current youth smokers who want to quit smoking.

```
• Read Data
• Filter Data to select only required columns
• Filter data to select only measures of smokers who want
  to quit
• Eliminate missing or NA values corresponding to smokers
  who want to quit
• Compute mean of percent of smokers who want to quit.
```

Figure 5. Pseudocode of Summary Statistics for Percent of Current Smokers who want to Quit

V. Conclusion And Future Research

From the study it is seen that there are more number of youth male smokers than youth female smokers. Smoking for a longer duration can increase damage to the smoker's body. Previous research has shown that if a smoker quits by age 30, the health of the smoker can become almost as good as a non-smoker [25].

The results in this study also stresses for researchers to come up with different measures for the youth smokers to quit smoking or to decreasing smoking. This can be done by spreading awareness among people such as through social media for instance [6].

Future research can be performed on the impacts of tobacco smoking on the health of the smokers. A detailed research on other factors related to smoking such as smokeless tobacco can also be performed.

References

[1] CDC (2016). Current Cigarette Smoking Among Adults in the United States. Retrieved on June 5, 2017, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

[2] Centers for Disease Control and Prevention (2016). Cigarette Smoking Among Adults—United States, 2005–2015. Morbidity and Mortality Weekly Report 2016;65(44):1205–11.

[3] Centers for Disease Control and Prevention (2017). Youth Tobacco Survey (YTS) Data. Retrieved on April, 2, 2017, from https://catalog.data.gov/dataset/youth-tobacco-survey-yts-data

[4] Danaei, G., Ding, EL., Mozaffarian, D., Taylor, B., Rehm, J., Murray, CJL., Ezzati, M. (2009). The Preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Med.

[5] Delnevo, CD., Bauer, UE. (2009). Monitoring the tobacco use epidemic III: The host: data sources and methodological challenges. Preventive Medicine;48(1 Suppl): S16–S23.

[6] Fiore, M., Croyle, R., Curry, S., Cutler, C. (2004). Preventing 3 Million Premature Deaths and Helping 5 Million Smokers Quit: A National Action Plan for Tobacco Cessation. American Journal of Public Health.

[7] Garfinkel, L. (1997). Trends in cigarette smoking in United States. Retrieved on June 25, 2017, from https://www.ncbi.nlm.nih.gov/pubmed/9245664

[8] Holbrook, JH., Grundy, SM., Hennekens, CH., Kannel, WB., Strong, JP. (1984). Cigarette smoking and cardiovascular diseases: A statement for health professionals by a task force appointed by the steering committee of the American Heart Association.

[9] Marr, B. (2015). How Big Data is Changing Healthcare. Forbes. Retrieved on June 22, 2017, from https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#38925f592873

[10] Mathers, C.D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, vol. 3, no. 11, pp. 2011–2030.

[11] Mokdad, AH., Marks, JS., Stroup, DF., Gerberding, JL. (2000). Actual causes of death in the United States. *JAMA* 2004, 291:1238–1245.

[12] Murray CL, Abraham J, et al (2013). The state of US health, 1990-2010: Burden of diseases, injuries, and risk factors. US Burden of Disease Collaborators, *JAMA*, 310:591–608.

[13] NBC News (2016). CDC Finds. Retrieved on May 22, 2017, from, http://www.nbcnews.com/health/health-news/only-15-percent-u-s-adults-now-smoke-cdc-finds-n579646

[14] NSDUH (2012). Prevalence of current smokeless tobacco use among adults 18 years of age and older, by selected characteristics. National Survey on Drug Use and Health.

[15] Office of the Surgeon General (US), Office on Smoking and Health (US) (2004). The Health Consequences of Smoking: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US). Reports of the Surgeon General.

[16] Public Health (2017). Smoking in America. Retrieved on June 2, 2017, from http://www.publichealth.org/public-awareness/smoking-in-america/

[17] R (2017). The R Project for Statistical Computing. Retrieved April 10, 2017, from https://www.r-project.org/

[18] Raghupathi W (2010). Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis. p211–223.

[19] Siahpush, M., McNeill, A., Hammond, D., Fong, G. (2006). Socioeconomic and country variations in knowledge of health risks of tobacco smoking and toxic constituents of smoke: results from the 2002 *International Tobacco Control (ITC) Four Country Survey. BMJ Journals.* Tobacco Control. Volume 15.

[20] Substance Abuse and Mental Health Services Administration (2012). The NSDUH Report: Substance Use Among 12th Grade Aged Youth by Dropout Status. Rockville (MD): Center for Behavioral Health Statistics and Quality.

[21] Tableau (2017). Tableau. Make your data make an impact. Retrieved on April 11, 2017, from https://www.tableau.com/

[22] Trivedi, S., Dey, S., Kumar, A., Panda, T. (2017). *Advanced Data Mining Techniques and Applications for Business Intelligence*. IGI Global. Information Science Reference. 400 pages.

[23] U.S. Department of Health and Human Services (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

[24] U.S. Department of Health and Human Services (2000). Reducing Tobacco Use. A Report of the Surgeon General. Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

[25] U.S. Department of Health and Human Services (2010). A Report of the Surgeon General: How Tobacco Smoke Causes Disease: What It Means to You. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.