

3/30/2017

Smoking Trends among the Youth of USA



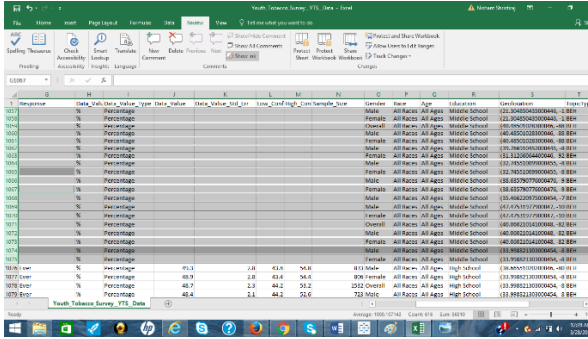
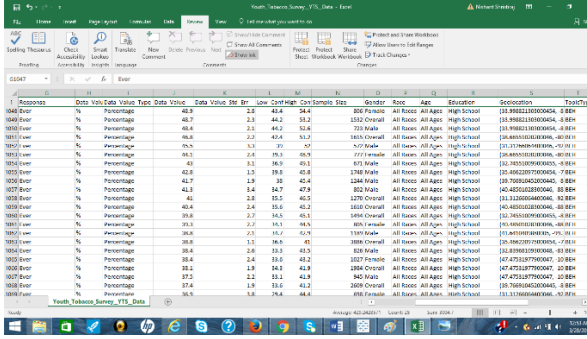
DATASET

Dataset URL: <https://data.cdc.gov/api/views/4juz-x2tp/rows.csv?accessType=DOWNLOAD>

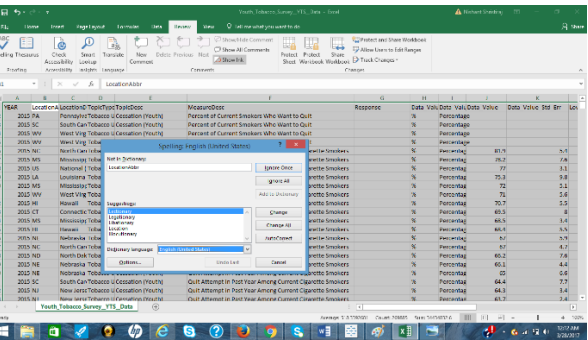
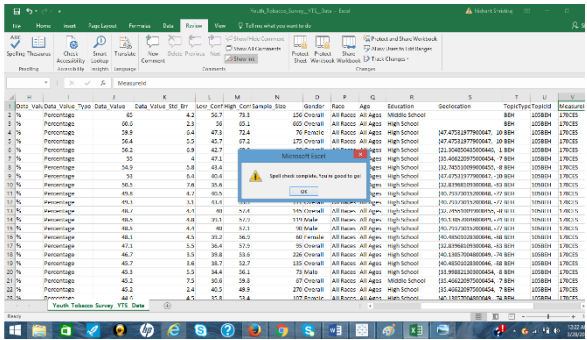
The Youth Tobacco Survey(YTS) provided the comprehensive dataset from 1999-2015 with over 20 columns and 9000 rows. There are multiple dimension topics such as cigarette smoking frequency and prevalence, user status in smokeless tobacco products from all the generations (1st Gen, 2nd Gen and 3rd Gen), percentage of smokers who are trying to quit and, the number of quit attempts in the past year among cigarette smokers to provide in depth analysis so the topic. The sample of the complete dataset has been the youth in the USA from middle and high schools across all the states. Different ages and races haven't been segregated which limits few of the insights such as tobacco use trends among people from different races of different age. At the same time, the topic description of cessation, cigarette smoking and smokeless tobacco gives us the option of filtering our data in the required fashion. The data seems to be combined from different sources by YTS, so 4 different stratifications have been used to bring them together. The values in the response tables is available only for the description of user status and smoking status. The topic description of Cessation has two parts: quit attempts in the past year by cigarette smokers and smokers who want to quit, which when used as proper filter can result in interesting insights.

DATA CLEANING

1. Missing Values

| Dirty Data | | Clean Data/ Remarks | |
|---|--|--|--|
|  | |  | |
| Empty values in the columns of data value, high confidence limit, low confidence limit and sample size for many rows. | | Subsequent rows with null values have been removed. | |

2. Misspellings

| Dirty Data | | Clean Data/ Remarks | |
|--|--|--|--|
|  | |  | |
| Running the spell check for the entire data to find misspelt words such as desc for description, abbr for abbreviation | | Corrected such errors. | |

3. Irrelevant Columns

Dirty Data

The screenshot shows an Excel spreadsheet with columns: Year, Location, Tobacco Use Type, Response, Data Value, Data Source, Data Value, Footnote, and Data Value. The 'Data source', 'Data Value', and 'Footnote' columns contain irrelevant information for the analysis.

Columns such as Data source and Data Value Footnote are not relevant to the analysis.

Clean Data/ Remarks

The screenshot shows the same Excel spreadsheet after removing the irrelevant columns. The columns are: Year, Location, Tobacco Use Type, Response, Data Value, and Data Value. The 'Data source', 'Data Value', and 'Footnote' columns have been deleted.

Columns(irrelevant) deleted from the dataset.

4. Duplicate Values

Dirty Data

The screenshot shows an Excel spreadsheet with columns: Year, Location, Tobacco Use Type, Response, Data Value, Data Source, Data Value, and Data Value. A 'Remove Duplicates' dialog box is open, indicating that there are duplicate values in the dataset.

Checking for duplicate values in the entire dataset.

Clean Data/ Remarks

The screenshot shows the same Excel spreadsheet after removing the duplicate values. The columns are: Year, Location, Tobacco Use Type, Response, Data Value, and Data Value. The 'Data source', 'Data Value', and 'Footnote' columns have been removed.

Duplicate values found (empty cells) and removed.

5. Irrelevant Data

Dirty Data

[illegible]

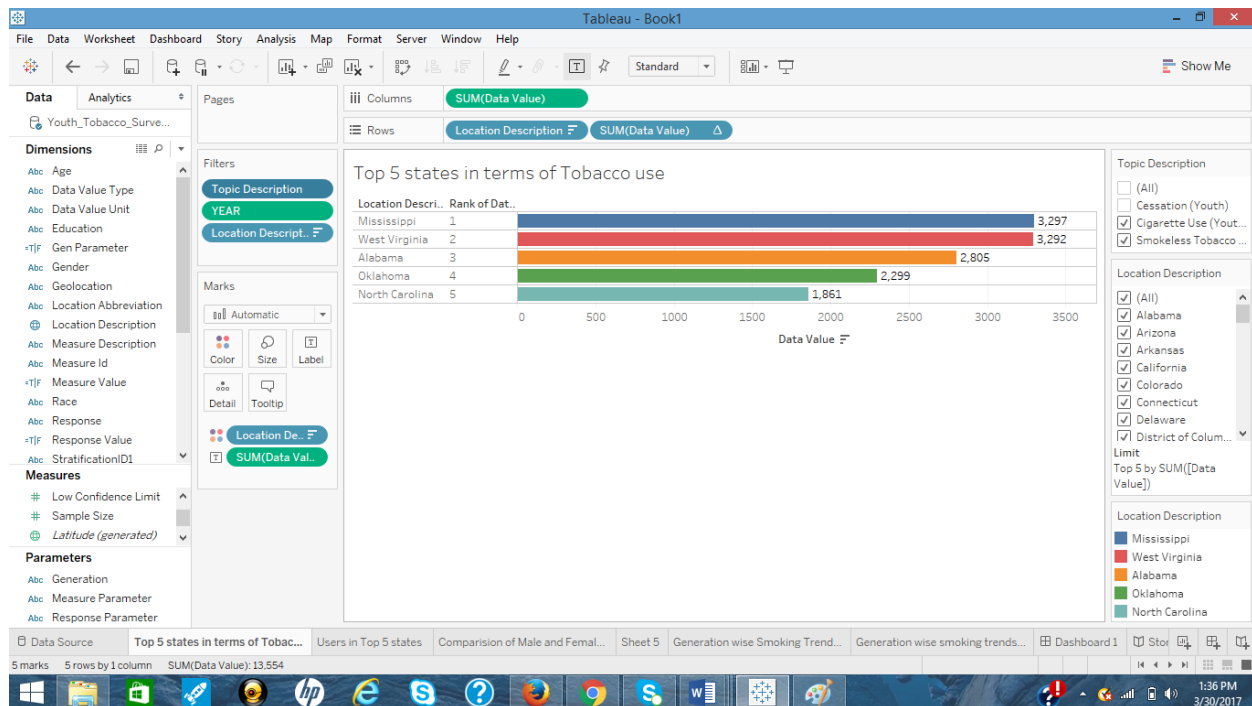
Data value field has percentage values which should lie within the range 0-100.

Clean Data/ Remarks

Proper filter range(0-100) applied and data has been validated.

DATA VISUALIZATION

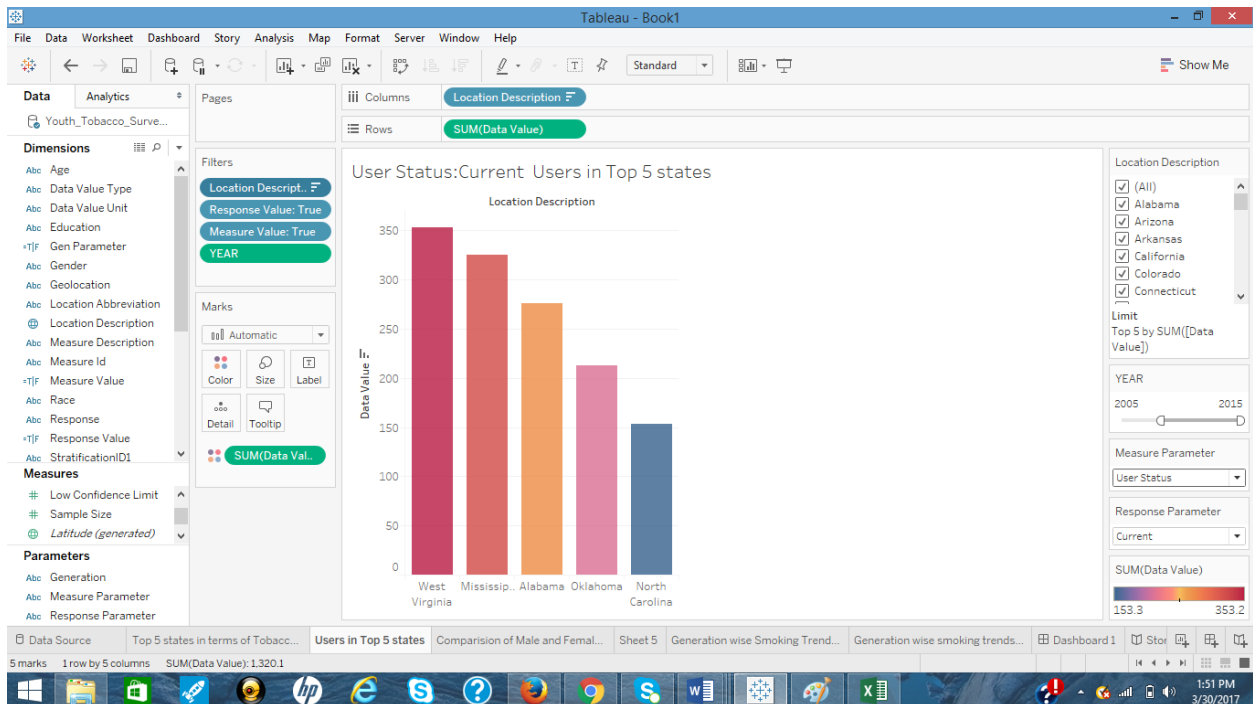
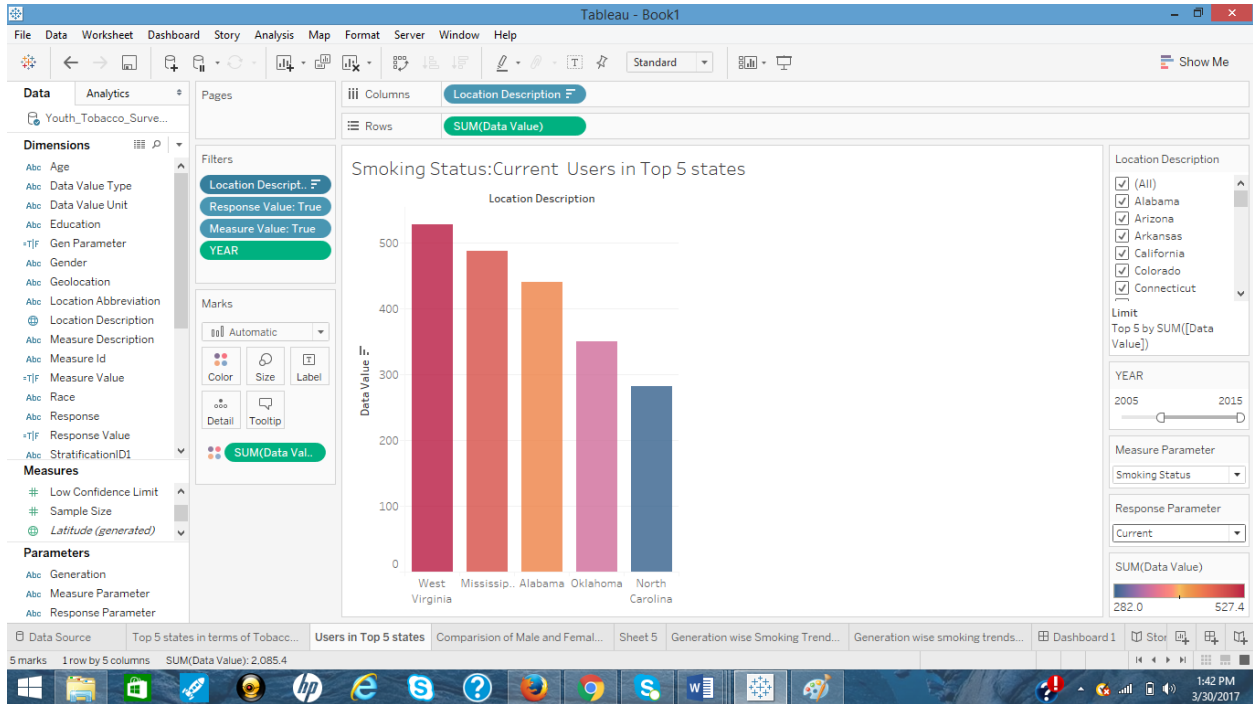
1. What are the top 5 states in terms of Tobacco products consumption (cigarettes and smokeless) among the youth of the USA?



[Visualization Tools Used: Rank]

The above horizontal bars show the Top 5 states with their rank and the number of tobacco consumers in each of these states in the USA from 2005-2015. The available dataset has values from the year 1999, but the year filter helps us to analyze recent data. We can clearly see that the state of Mississippi has the maximum tobacco consumers among the youth from high school or middle school. The available value in the dataset is also recorded for cessation of cigarette smoking, which in case of this visualization has been filtered out, as we need the current numbers of smokers in these states. The ranking of these states has been obtained based on the sum of the values in all the states and extracting the top 5 states thereafter.

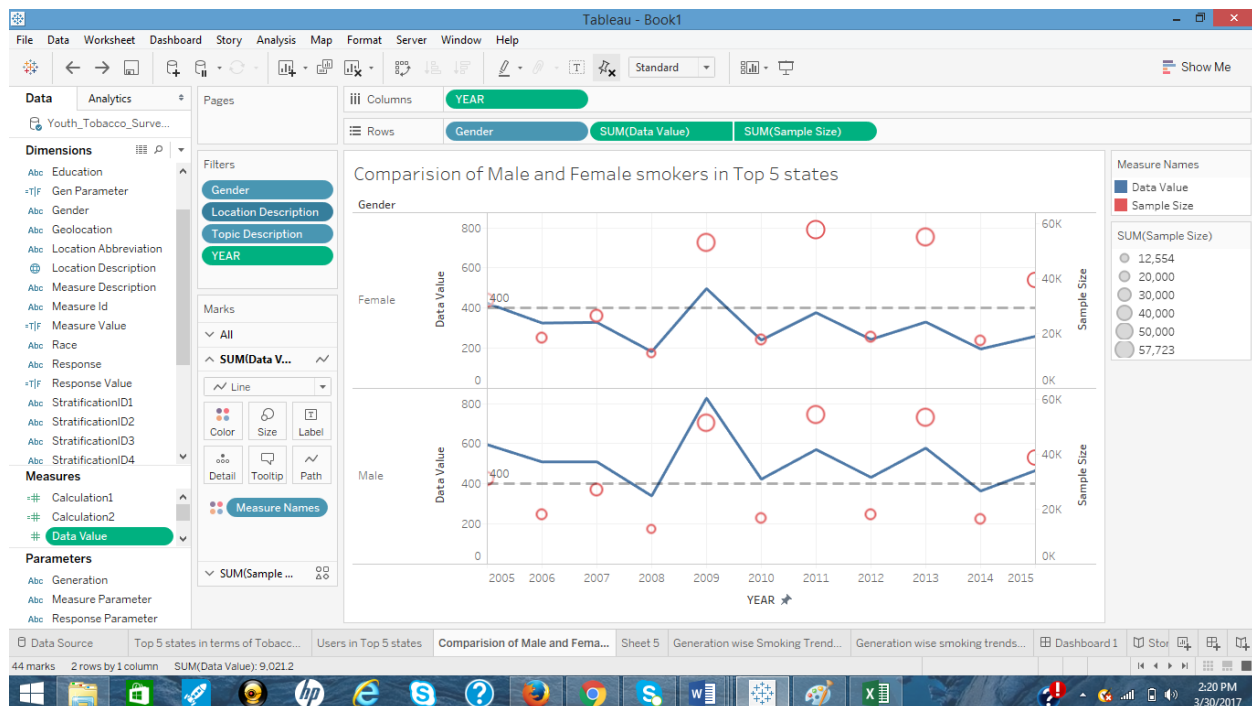
2. What are the cigarette smoking prevalence and frequency, and smokeless tobacco prevalence among the youth in the Top 5 states?



[Visualization Tools Used: Parameters and Calculated Fields]

From the above visualization, two different parameters named ‘Measure Parameter’ and ‘Response Parameter’ can be noticed. The measure parameter is having two values: smoking status and user status. Smoking status refers to the cigarette smoking among the youth and user status corresponds to the smokeless tobacco use. The response to both these parameters can be either current, frequent, or ever. ‘Measure value’ and ‘Response value’ are two different calculated fields created to map the parameters to the respective dimensions. The response to the smoking status shows that from 2005-2015, West Virginia has the highest number of current cigarette smokers and North Carolina has the lowest among the youth from all education levels. Similarly, in terms of smokeless tobacco use, West Virginia ranks the highest.

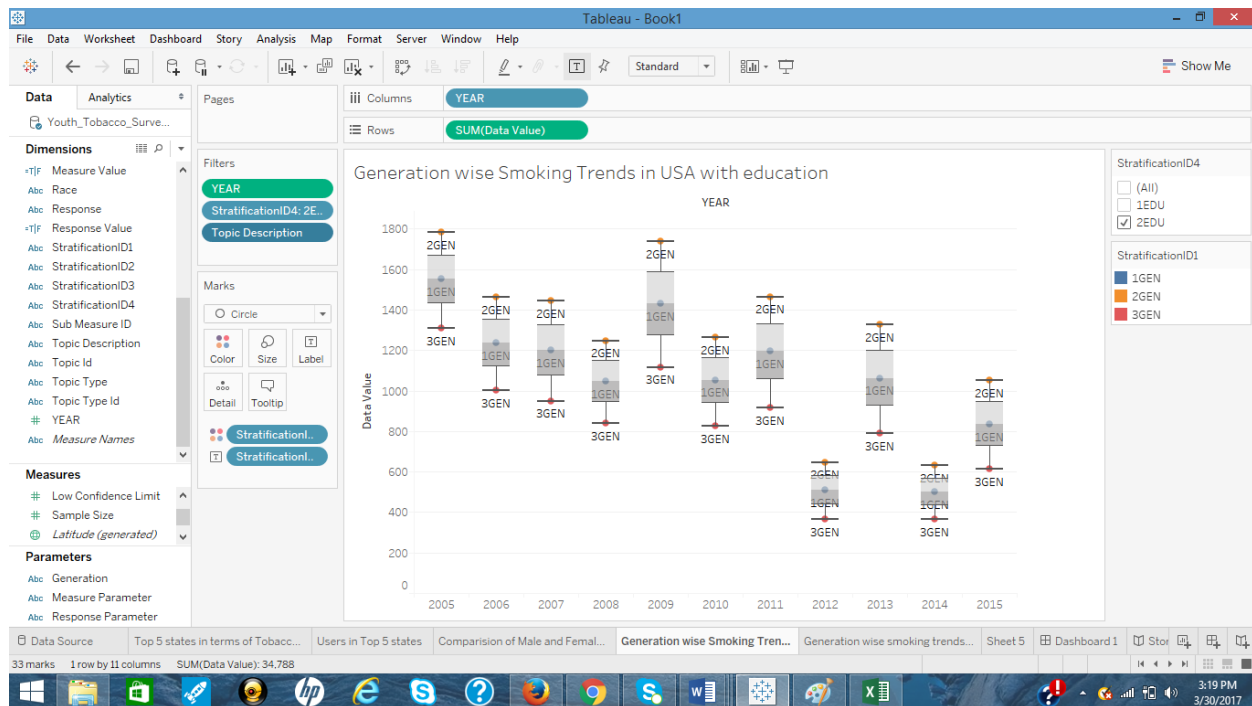
3. What is the comparison between males and females in terms of tobacco use in the Top 5 states?



[Visualization Tools Used: Dates, Dual-Axis Chart, and Reference Line]

The line graph shows the number of users (males and females) combined in the top 5 states. The circle and its size determines the size of sample considered for the study. The result is filtered for cigarette and smokeless tobacco users only from the year 2005-2015. Clearly, number of male smokers in the top 5 states combined is higher than the number of female smokers every year. The number of females is always below the reference line which, apart from the year 2009, where the spike is just above it. There is a significant increase in the number of male smokers too in 2009, and then dropping in 2010. The number of male smokers have mostly been above the reference line apart from 2008 and 2014.

4. Which generation (1st Gen, 2nd Gen, 3rd Gen) of smokers consumed tobacco the most?

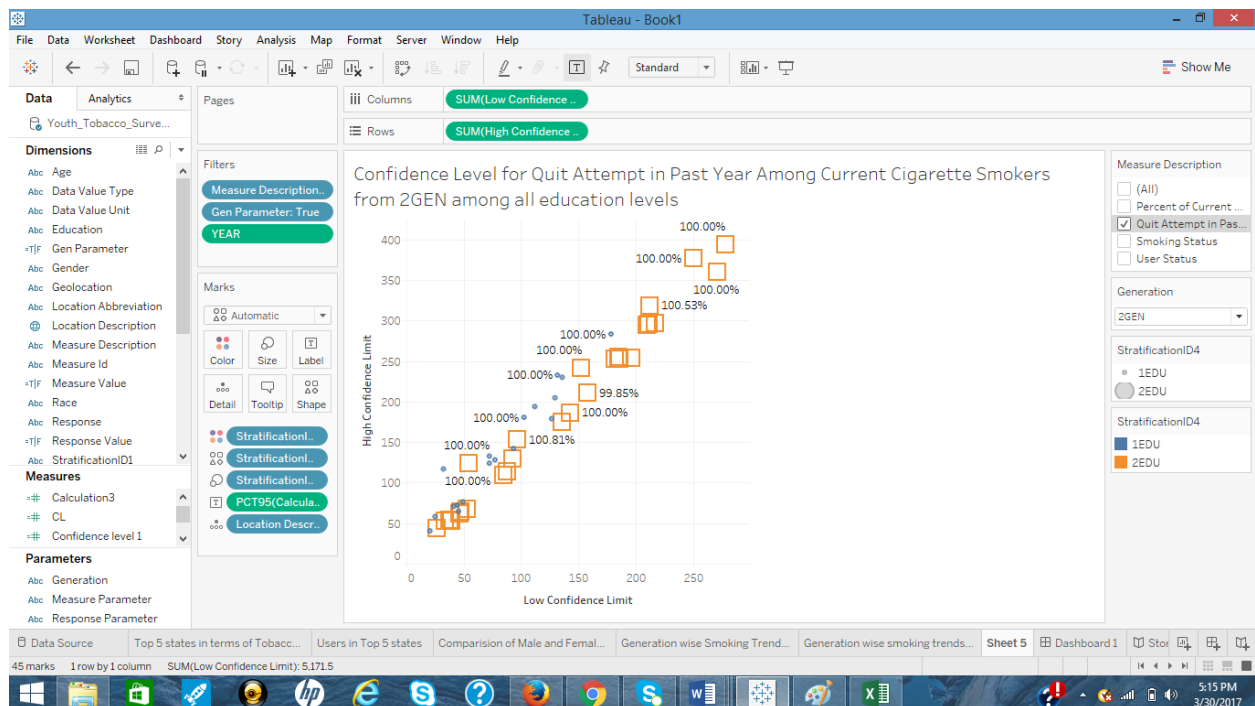


[Visualization Tools Used: Box and Whisker Plot]

By carefully observing the Box and Whisker plot, the 2nd Generation smokers are the ones who consumed the tobacco most. Also, the people from high school are larger in number in terms of

tobacco consumption than the ones at middle school. The 3rd Generation smokers seems to have less affinity for tobacco products. The education here is also indicative of age since the ones studying in high school will be elder than the middle school youth. So, we can bring forth the point that age also contributes to tobacco consumption. A simple prediction can be made with the increased number of tobacco consumers at high school that a person who consumed tobacco at middle school is more likely to continue doing that at high school and he will be joined by new people who start tobacco smoking at high school. The number of smokers in even years also tend to be lower than the odd year before it. So, the number of smokers may go down in 2016 for all generations as compared to 2015.

5. What are the confidence levels for different measure descriptions among the youth of the USA for different generations?



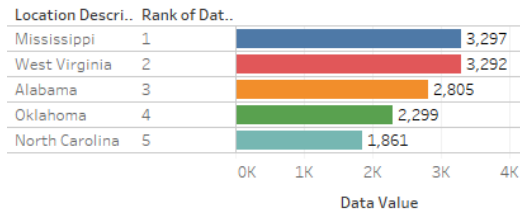
[Visualization Tools Used: Scatter Plot, Calculated Fields, and Parameters]

The confidence level analysis will give an estimate of accuracy of our dataset for different measure description. Confidence limits are usually used when we have a sample, we carry out our analysis on that sample and try to estimate the results for the whole population [1]. The closeness of the estimated result to the real result is the confidence intervals. In our case the confidence level should ideally be 100% as all the samples in the dataset are individual samples and there is no estimation done. A parameter Generation is used to analyze data for different generations. From the visualization, the actual result shows a confidence level of more than 98% in most of the cases, and we can assume the dataset to be relevant and authentic. The calculation for the confidence level can be done as: $[(\text{High Confidence Limit} - \text{Data Value}) / (\text{Data Value} - \text{Low Confidence Limit})]$.

DASHBOARD

Smoking Trends among Youth of the USA: A Methodical Approach

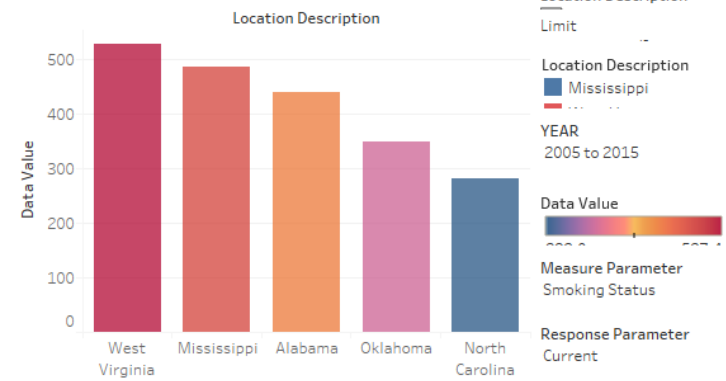
Top 5 states in terms of Tobacco use



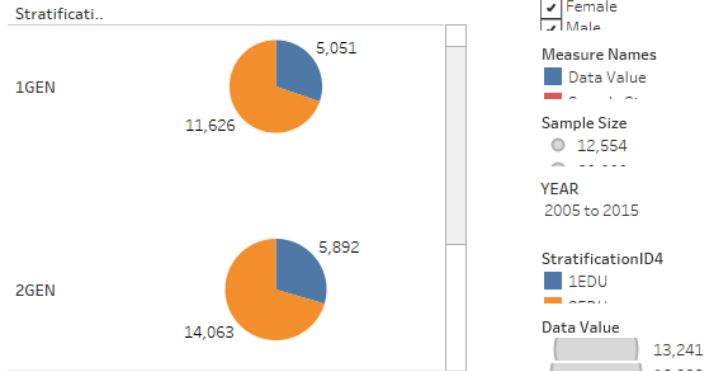
Comparison of Male and Female smokers in Top 5 states



Smoking Status: Current Users in Top 5 states

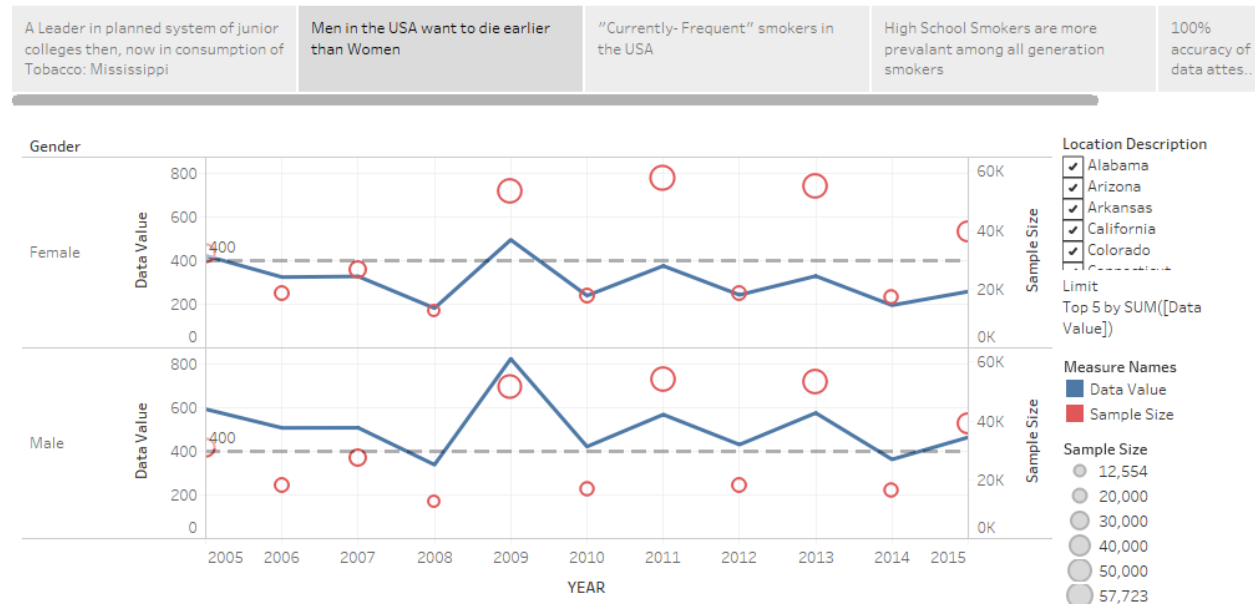


Generation wise smoking trends in USA from 2005-2015



STORY TELLING

Smoking Trend Model of the Youth in the United States

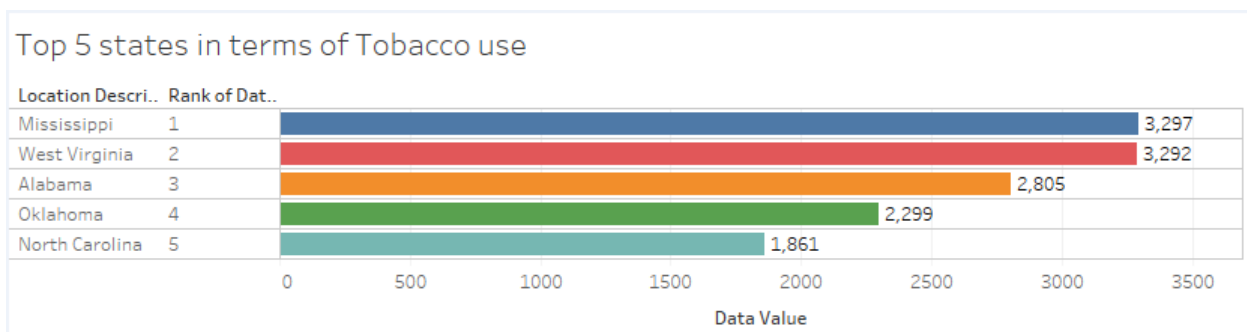


In 2000, a student from the Graduate School of Management, University of California-Irvine, carried a case study and found that anti-smoking campaigns are most cost effective and significantly helped in reducing adolescent smoking prevalence at a low per capita income among the youth between 12-16 years [2]. The adolescent smoking problems in USA has been prevalent from past few decades. Government agencies carry out regular surveys to analyze and end the root cause of this internal epidemic and prevent the youth in USA from chronic diseases. In a report published by Centers of Disease Control and Prevention, 5.6 million citizens are expected to die prematurely with the current smoking rate [3].

Not only in the USA, the problem of smoking leading to various chronic diseases, is existent in many other parts of the world. Every major country in the world realizes the potential of its youth and wants protect them, gives them quality education, and prevent them from any negative

influences so that they can serve their nation in need. As I smoker who has quit smoking, this topic is ideal for me to analyze as a part of my academic project. My driving facts for this study were few stats, one of them being 25.3% high school students were reported to be current user of any tobacco products and 13% of them used more than 2 different products [3].

In the beginning, I wanted to narrow down my analysis study from a comprehensive dataset from Youth Tobacco survey, which gave data from 1999-2015 for all the states of the USA. All my analysis is for the stretch of 10 years, i.e., 2005-2015 to get a more recent result. I found out that Mississippi and West Virginia are having high numbers of individuals who are addicted to use tobacco products currently, as shown below. The data to get these states was filtered by considering



the people who either use cigarette or smokeless tobacco products. The cases and attempts of cessation was excluded for this finding. The next finding was interesting to note because the sample considered was not uniform for this analysis, still men were rated as higher consumers of tobacco products than women in these states. The ratio of the data value with corresponding sample was used to get the result. It is estimated worldwide that men nearly smoke 5 times more than women [4]. The line chart visualization also showed that the number of users (men and women) tend to be more in the odd year than the next even year. I also observed different measures such as smoking status and user status in these 5 states. West Virginia is always having more number

of users in terms of current, frequent, and ever users of smokeless tobacco products. For cigarette smoking, Mississippi is more prone to ever smokers than West Virginia.

Smoking can develop as a habit in an individual at an early age, resulting from various external or internal factors. External factors include advertisement, movies which has scenes related to smoking, friends, etc. One of the most evident internal factors of smoking could be if some member(s) of the family also smoke(s). In the top 5 states of the USA, I observed that 2nd Generation smokers are more in number and 3rd Generation is the least. 1st Generation smokers lie in the middle 50%. When compared with the education level, youth in the middle school were low in numbers than the youth in high schools. Now this is an interesting fact as school health programs are constantly being implemented in all parts of the USA from middle schools and even parents are being involved in these programs to prevent tobacco use by kids [5]. Still after all these precaution, the number of smokers goes up in high school. This explains the loopholes in the measures taken by these school health programs. Also about the increase in number of smokers, we can estimate that the number of smokers who already existed in middle school continue to smoke in high school.

Smoking as a habit can't be left overnight. It is very difficult to leave something to which one has been habituated for years. Government in the US has been taking several measures to tackle the problem and bring out their youth from this curse. Per my analysis, these measures should be more centric to states of Mississippi, West Virginia, Alabama, Oklahoma, and North Carolina as these states have higher number of tobacco consuming youth. The Centers for Disease Control and Prevention(CDC) and NGO's should work towards developing a centralized model to eradicate smoking from the youth and test that model at first in these 5 states, which later can be implemented throughout the country.

Quitting to smoke is something which I have achieved on a personal level. All the efforts and methods used to leave smoking will go in vain without proper motivation, support, and a strong will power to leave smoking conclusively.

REFERENCES

- [1] "Confidence interval." *Wikipedia*. Wikimedia Foundation, 29 Mar. 2017. Web. 30 Mar. 2017.
- [2] Pechmann, C. "Anti-smoking advertising campaigns targeting youth: case studies from USA and Canada." *Tobacco Control* 9.90002 (2000): 18ii-31. Web.
- [3] "Morbidity and Mortality Weekly Report (MMWR)." *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, 14 Apr. 2016. Web. 30 Mar. 2017.
- [4] "Gender empowerment and female-to-male smoking prevalence ratios." *WHO*. World Health Organization, n.d. Web. 31 Mar. 2017.
- [5] "Guidelines for School Health Programs to Prevent Tobacco Use and Addiction." *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, n.d. Web. 31 Mar. 2017.