Factors Affecting Tobacco Use Among Young People in the US

CSE 163 Final Team Project

Guicheng Ma, Yuqing Li

Summary

1. How did the tobacco use rate change over years?
    a. Tobacco use in the United States is generally decreasing.
2. Does the Tobacco Legislation in each state affect young people's tobacco use habits?
    a. The Tobacco Legislation does affect young people's tobacco use habits.
3. Are there factors (such as race or gender) other than the legislation affecting young people's tobacco use? Based on the information of a young person, can we use that information to predict the possibility of a young person smoking?
    a. By analyzing the data we have, there is no explicit relationship between other factors and tobacco use rate. Therefore no prediction could be made
4. What are the trends in cigarette and e-cigarette popularity? Should the direction of future legislative rules be improved or the emphasis be changed according to this popular trend?
    a. Cigarette use fell sharply, smokeless tobacco use fell slightly, but e-cigarette use rose sharply.
5. Apart from the current factors we knew from the datasets, what other possible factors might be related to the smoking rate? How could we gather that information?
    a. We gather more data by making a survey in the terminal that takes more information on other factors and stores them in a CSV file. And give users information that could help them quit smoking.

Motivation

Smoking leads to disease and disability and harms nearly every organ of the body. For young people, tobacco is even more harmful. Unfortunately, tobacco contains nicotine, an ingredient that can lead to addiction, so people will find it difficult to quit. Most of them keep using tobacco for their whole lives. Every year, tobacco causes more than 480,000 deaths in the United States, including more than 41,000 deaths resulting from secondhand smoke exposure. This is about one in five deaths annually, or 1,300 deaths every day. Despite these well-known dangers, we often see young people around us smoking. For the sake of their health and public health, we need to improve laws to reduce tobacco use. The good news is that over the years, smoking rates have been reduced across the United States. We want to use the data we have to analyze what are the factors of this reduction and how we can improve these methods and provide better protection to them.

Dataset Description

1. CDC STATE System Tobacco Legislation - Youth Access

https://catalog.data.gov/dataset/cdc-state-system-tobacco-legislation-youth-access

This dataset, published by the Centers for Disease Control and Prevention (CDC), includes the legislation regarding youth smoking and tobacco sales in each state from 1995 to 2022.

2. Behavioral Risk Factor Data: Tobacco Use (2011 to present)

   https://catalog.data.gov/dataset/behavioral-risk-factor-data-tobacco-use-2011-to-present

   This dataset is published by the American Centers for Disease Control and Prevention. And it includes information on tobacco use prevalence by demographics, type of tobacco (such as cigarettes and e-cigarettes), use frequency, quit attempts, gender, age, and education from 2011 to 2019.

Method

1. How has the tobacco use rate changed these years?

   This is a new question we raised after the proposal. In order to better show the change in rate, we decided to make an image that could change with the timeline. We found a library pyecharts that could implement this function. We groupby the data according to time and location to find out the sum of various tobacco usage rates in different years in each place. Represent these data in an image that varies with the time axis. The output of the image is in the form of an HTML file. In this map, We can see the change in smoke rate in the US by clicking the start button and see the exact rate of a state in a year by clicking on it.

2. Does the Tobacco Legislation in each state affect young people's tobacco use habits?

   Before we start writing codes to investigate our research questions, we need to pre-process the datasets. For the CDC STATE System Tobacco Legislation dataset, since the legislation does not change very often, we want to only keep the data on the fourth quarter of each year in each state. So we will filter out the data we want to keep and get rid of the rest of them. Since the Behavioral Risk Factor Data: Tobacco Use (2011 to present) dataset only contains the data from 2011 to 2019, in order to match with the time range, we also want to just save the data from 2011 to 2019 in the CDC STATE System Tobacco Legislation dataset. For the Behavioral Risk Factor Data: Tobacco Use (2011 to present) dataset, since we are investigating the tobacco use of young people, we could

just filter the data (age) to by '18 to 24 Years' and only keep those. And we also want to filter the ['MeasureDesc'] column, which represents the smoking status of people, to only keep values equal to 'Current Smoking' and 'Current Use'. For both datasets, we want to get rid of the unnecessary column of data (those that are not related to our research question) such as the TopicId.

This question is to find out whether changes in legislation are associated with changes in tobacco use rates. We first find the total percentage of young people (under 24 years old) who use tobacco for each state in a given time frame and each year. I finish this step by using ‘groupby(['LocationDesc', 'YEAR'])['Data_Value'].sum()’. Then I subtract the percentage of 2011 from the percentage of 2019 to calculate the net change of each state in the time range. Then I find the names of the three states with the most drops in net change and the names of the three states with the most rises (or including the least drops) in net changes. I do it by using sort_values to sort the data and using .head(3) and .tail(3) to get the exact rows. I then look at changes in legislation in those six states. I then look at the time data under the effective date column (which is when the law went into effect) to find states that had legislative changes within the time frame. I will then make a graph of the changes in the total tobacco use rate in these states over the past few years and compare it with the effective date of new legislation in each state to see if the use rate has changed after the legislative change.

3. Are there factors (such as race or gender) other than the legislation affecting young people’s tobacco use? Based on the information of a young person, can we use that information to predict the possibility of a young person smoking?

   We built a machine learning model to make this prediction on the percentage of young people using tobacco, based on the year, location, which kind of tobacco they are using, gender, race, and education level. In order to find the best predicting model, we tried to change the hyperparameter. We tried different ratios of train sets and test sets. Also, we tried different depths to avoid overfitting. If we could find a test size and depth that is good enough with small error, then we will see if there is any relationship between these factors and the tobacco use rate and also make a prediction on the possibility of a young person with information including location, gender, race and so on smoking. By predicting the possibility of a young person smoking and comparing the possibility with the overall situation in the US, we could see where the help of smoke quitting or education on tobacco is most needed(where the proportion of certain kinds of young people is high)

4. What are the trends in cigarette and e-cigarette popularity? Should the direction of future legislative rules be improved or the emphasis be changed according to this popular trend?

   In order to answer this question, I want to create a line graph that will contain three lines that show the change in the percentage use of each kind of tobacco over years. I use groupby(['TopicDesc', 'YEAR'])['Data_Value'].sum() to sum up the percentage use of each tobacco product in each year in the US. Then in order to find the average percentage in that year, I divide the sum of the percentage by the number of states in each year. However, there are only 33 states that have data for E-Cigarette use in 2018, so I only divide that sum by 33. Then I plot all the data into a plot. The x-axis is the year, and the y-axis is the percentage of each kind of tobacco use.

5. Apart from the current factors we knew from the datasets, what other possible factors might be related to the smoking rate? How could we gather that information?

   We built an interactive window in the terminal that could give users a survey and gather the information provided with permission in a CSV file. We could use the information gathered from the survey to help with future analysis. Also, we want to give users useful suggestions and information that could help them quit smoking. But since we don't have enough time to collect information that could be provided. So we leave the suggestion part as a future goal.

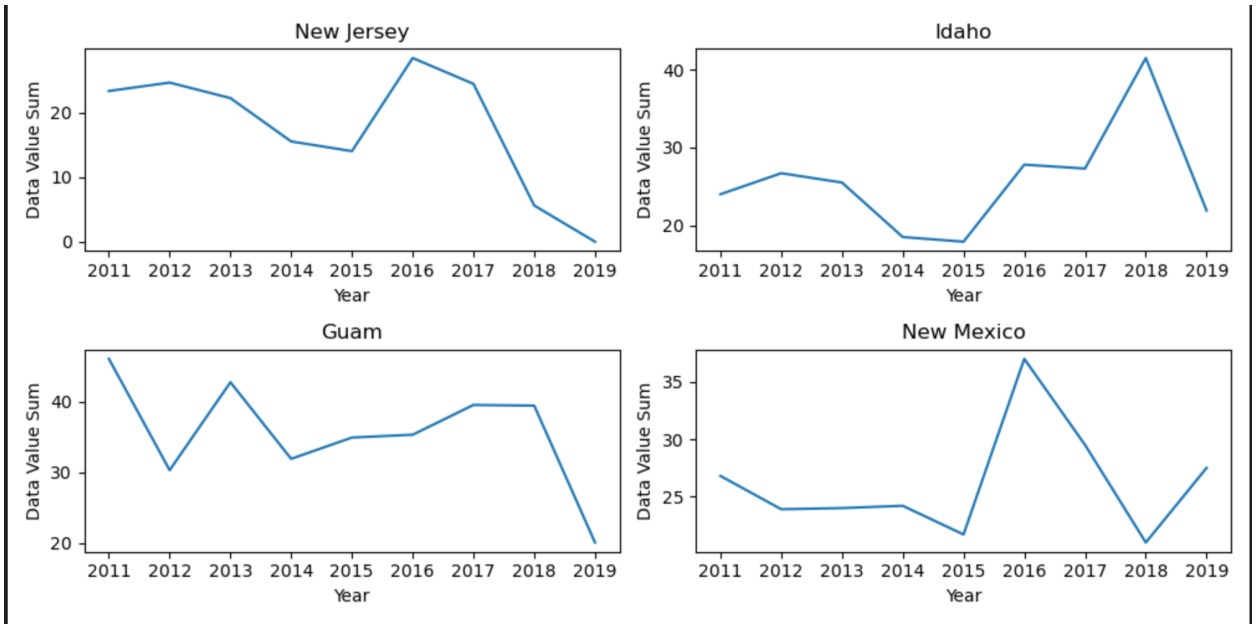
Result

1. We can see the overall tobacco use rate is decreasing since 2011. But there is a short rise in 2016 and 2017. It is because the E-cigarette use data was collected and turned the general rate higher than before.
2. The answer to this question is there is the total percentage change in tobacco use is related to legislation change. Because I could see the top 3 states with the most decrease in net change and the top 3 states with an increase or the least decrease in net change are those:

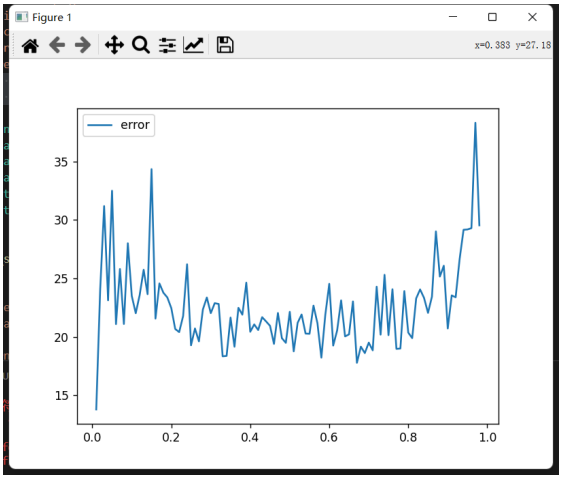
   By checking the time of legislation change date in those states in the graph I created, I find out that after the year of new legislation change, the percentage of tobacco use also change. For instance, in New Jersey, the new legislation changed in 2017, and there is a huge decrease in percentage from 2017 to 2018, this could prove that legislation could actually change the percentage of tobacco use. Another interesting finding is that some

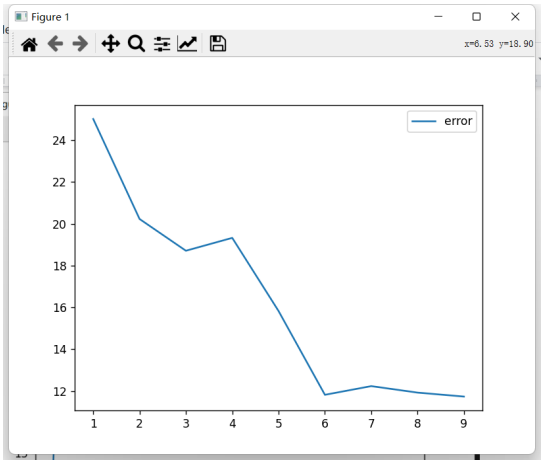
legislation changes can also lead to an increase in smoking rates. I think this may indicate that the change is ineffective or affected by other factors.

{'Idaho': [Timestamp('2015-06-01 00:00:00')], 'New Mexico': [Timestamp('2015-06-19 00:00:00')], 'Guam': [Timestamp('2014-05-21 00:00:00'), Timestamp('2018-01-01 00:00:00')], 'New Jersey': [Timestamp('2017-11-01 00:00:00')]}



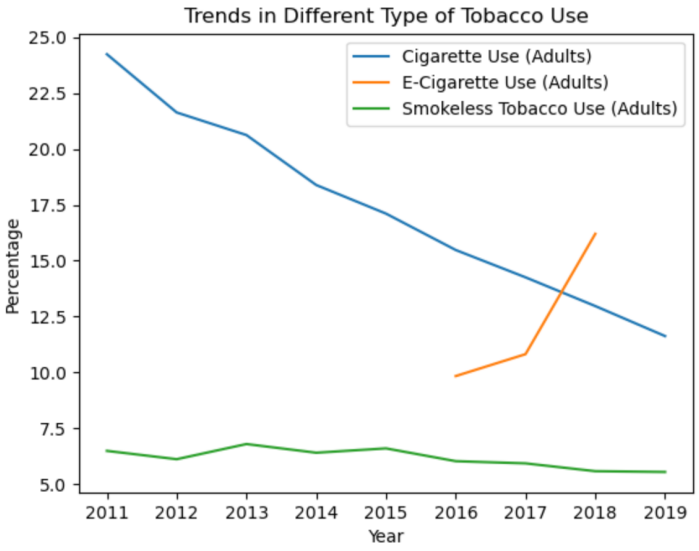
3. There is no apparent relationship between other factors and the tobacco use rate. We tried to change different hyperparameters to build a model that makes the best prediction. By predicting the possibility of a young person smoking and comparing the possibility with the overall situation in the US to decide if we need to pay more attention. However, even though the result seems somewhat reasonable when we run it



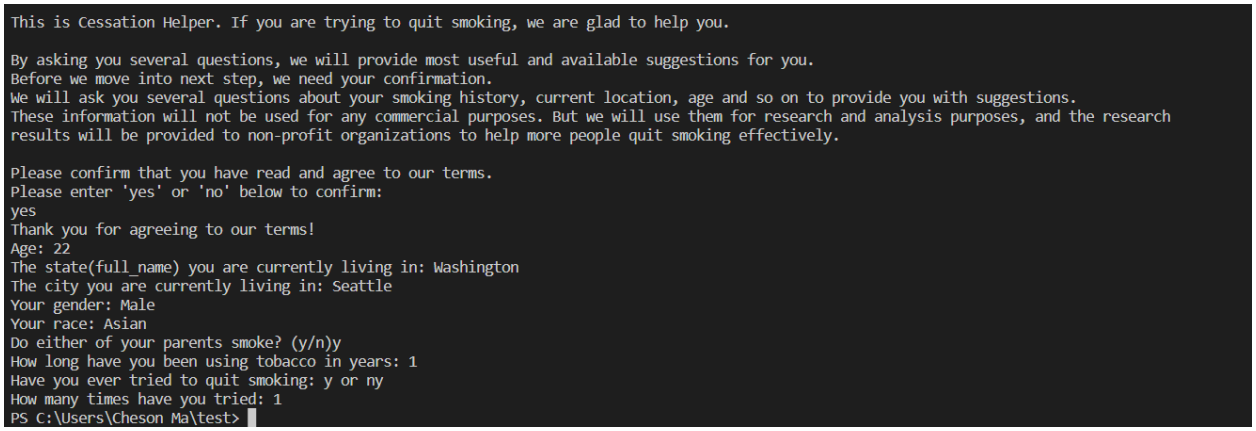


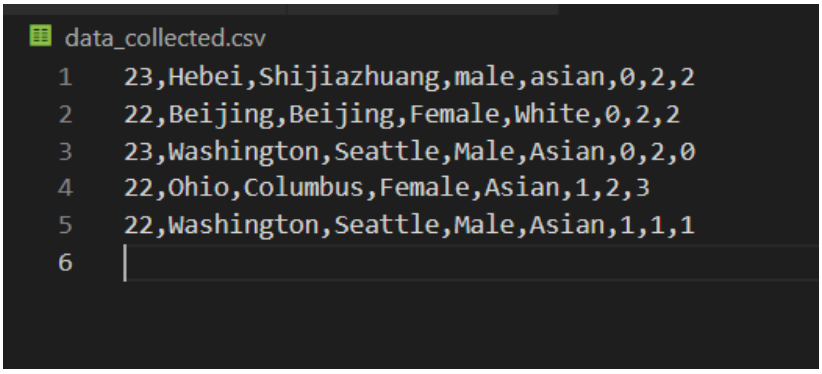
The thing is that when we run it again, the result will be totally different without any relationship. So we got the result that the Machine Learning technique, at least the Decision Tree model, is not quite stable. We don't think it can provide us with any analyzable information or reasonable prediction.

4. The answer to this question is that the use of cigarettes has fallen substantially, and the use of smokeless tobacco has declined slightly, but the use of e-cigarettes has risen substantially. This trend clearly illustrates the changing choice of tobacco types among young people. Since the previous question confirms that legislation will have an impact on tobacco use, the government should follow the selection trend of tobacco types and produce more restrictions and penalties on e-cigarettes in subsequent legislation, especially for young people.



5. We provide users a survey in the terminal that could gather the information users input. For ethnicity, we asked permission to use this information at the beginning of the survey. And we also want to provide users with useful and available suggestions. But it takes a long time to gather all the useful information in the US. So we skipped this part and left it as future work.





Impact & Limitations

What are the potential implications of your findings? Who would benefit from your analysis and who would be excluded, or worse, harmed? Are there any biases in your data that could affect the results? Clearly outline your limitations of the analysis, and how others should or should not use your conclusions.

The potential impact of my research results is that it may affect the legislative focus of each state, allowing the government to consider more factors that may affect young people's use of tobacco and formulate relevant laws. Young people and policymakers will benefit from my analysis. Because my research can provide policymakers with the direction and focus of the legislation, so as to reduce the smoking rate of young people and maintain the health of young people. People of other ages using

tobacco are excluded but not harmed. A possible bias in my data is that the effective date of some government laws is empty, so we don't know if it was generated during the time we want to explore. In addition, the data of legislation is quite messy, so we cannot clearly analyze each legal clause to see which clauses may be more effective. For the machine learning part, although we have analyzed that factors such as race, education, etc. cannot help us predict the smoking rate. There is no apparent relationship shown. We hope to collect more data to increase our collection of data on possible influencing factors for further research. Others should use our conclusions, at least here we proved the relevance of legislation and tobacco use. Others can also use the interactive tool we only do to collect data to help them collect data for research.

Challenge Goal

1. Challenge Goals Same as Proposal, updated with any changes made during implementation. If the challenge goals were scaled back or expanded, explain why the task turned out differently than initially estimated.
    - Messy Data:
        i. The datasets we found were quite messy. We spent a long time trying to understand their meaning. It also takes a lot of time to think about what values we need and filter our dataset. We planned to join two data sets together, but we found it difficult to merge them. Instead, we could simply analyze them separately.
    - Machine Learning:
        i. We planned to use machine learning to see the relationship to predict how the tobacco use will change if the law is amended. However, we found the legislation in these years has not apparently changed, so it is unnecessary to make an attempt to build a machine-learning model based on the legislation. Instead, we made a machine learning model to predict the tobacco use rate based on people’s information of race, gender, location, and so on,
    - New Library:
        i. We used a new library pyecharts. It seems like a new library which is hard to find instructions for. But it is useful to build an interactive map that could better express the change in years
    - Interaction with a user for data collecting
        i. We made an interactive terminal that could take users' answers and store them in a new CSV file which could be used for future analysis.

Work Plan Evaluation

The work plan that we proposed was not quite accurate. Although we all agreed to start the project early, we only started two weeks before the project's due day. That was because both of us had busy schedules and we couldn't decide on the topic of the project because we couldn't find a suitable dataset. However, in the last two weeks, we have implemented the work plan well.

Testing Description

We tried to filter the dataset we have to a smaller data set. By comparing only six states instead of all states in the US and get the largest net-change ones and smallest net-change ones. We could see if our function finding the three largest ones and three smallest ones is right or not.

Collaboration

We used Google to search for the information or syntax we needed. And also some Youtube videos for help. For instructions of libraries, we got syntaxes from the official website.