The next step in the interview process is to complete the case study.

Please take a look at the survey and data available here:
https://www.cdc.gov/brfss/annual_data/annual_2017.html

This is the 2017 edition of the CDC's annual behavioral survey on risk factors. It can be used to understand smoking and e-cigarette usage in the US population. Please analyze the data to find the prevalence and demographics of smokers and e-cigarette users, and compare and contrast findings between these groups to identify meaningful differences.

Please create a presentation - could be a brief write-up or set of slides with visualizations - based on these findings and any other analyses you choose to do with these data. This presentation should also summarize any findings you think are particularly useful for describing the population of adult e-cigarette users.

Please submit your presentation and R/Python/other code used to analyze the data. The ease with which the reviewer can understand and review your code and approach will also be taken into account.

You have 7 days to complete the assignment and send it back to me.
Please let me know if you have any questions

Author: Pengcheng Wu
Submit date: Feb 19th, 2020

Ⅰ Overall Population Analysis

The questionnaires collect 450,016 effective respondents' answers in total. 358 variables are involved in this survey, which contain aspects from demographics, health status, social behaviors, etc. Considering the aim of the case study is to find and compare the differences between smokers and e-cigarettes users, we focus on splitting the respondents into two categories, smokers and e-cigarettes users by their answers in column ['SMOKDAY2'] and column ['ECIGNOW']. Before splitting the dataset, unnecessary variables, such as column ['DLYOTHER'], are removed because of the tricky data type. There are also many blanks in the dataset, representing the person was not asked or the answer is missing. We choose to replace blanks with 0 to alleviate the problems that may be caused by blanks.

Ignoring blanks in column ['SMOKDAY2'] and column ['ECIGNOW'], there are five categories in each of the variables, representing people's smoking frequency in daily life. According to the calculation, there are 63,466 people claiming they are currently smoking cigarettes, occupying 34.21% of the whole number of people who answered this question. 13,677 people claim they are currently using e-cigarettes, occupying 20.50% of the whole number of people who answered this question. Among the 63,466 smokers, 44,617 claim they smoke every day, taking 70.30% of the group. While in the 13,677 e-cigarettes users, 5,019 people claim they use e-cigarettes or other electronic vaping products every day, taking 36.70% of the group. The table below shows a summary of numbers of people who may have multiple smoking preferences.

Table. 1 Number of Smokers and E-cigarettes Users by Cross Statistics

	Smoke (every day)	Smoke (some days)	Smoke (not at all)	Col_sum
E-cigarettes (every day)	831	713	2950	4494
E-cigarettes (some days)	4014	1643	1424	7081
E-cigarettes (not at all)	20563	6848	13256	40667
Row_sum	25408	9204	17630	52242

According to the observation of two groups of people with different smoking preferences, we found the phenomenon of smoking and using e-cigarettes at the same time exist commonly among the respondents. Table.2 and table.3 show the percentages of people in every single situation. From the table, we found only 3.27% of daily smokers use e-cigarettes every day, while 18.49 daily e-cigarettes users smoke every day. If a person only smokes some days, it is very likely this person doesn't use e-cigarettes at all. However, if a person uses -e-cigarettes some days, it is very possible this person also smokes daily or some days. These conclusions reveal the relationship between different groups of smokers and e-cigarettes users, at the meanwhile, show the switching trend of them from tradition cigarettes to new electronic cigarette products.

Table. 2 Percentage of People in Different Situations (based on smoking frequency)

	Smoke (every day)	Smoke (some days)	Smoke (not at all)	Col_sum
E-cigarettes (every day)	3.27%	7.75%	16.73%	8.60%
E-cigarettes (some days)	15.80%	17.85%	8.08%	13.55%
E-cigarettes (not at all)	80.93%	74.40%	75.19%	77.84%
Row_sum	100.00%	100.00%	100.00%	100.00%

Table. 3 Percentage of People in Different Situations (based on e-cigarettes frequency)

	Smoke (every day)	Smoke (some days)	Smoke (not at all)	Col_sum
E-cigarettes (every day)	18.49%	15.87%	65.64%	100.00%
E-cigarettes (some days)	56.69%	23.20%	20.11%	100.00%
E-cigarettes (not at all)	50.56%	16.84%	32.60%	100.00%
Row_sum	48.64%	17.62%	33.75%	100.00%

Ⅱ Visualization of Demographics

Based on the result obtained before, we understand people may use traditional cigarettes and e-cigarettes at the same time. The outcomes of percentage statistics show 96.73% of daily smokers don't use e-cigarettes every day. 81.51% of daily e-cigarettes users don't smoke every day. These huge occupations of percentages make it possible for us to split the two specific groups mentioned above as the groups of smokers and e-cigarettes users for data visualization analysis. The huge number of entries in both groups guarantee the reliability of our conclusions in this section.

Before visualizing the distributions of the demographic variables, we need to completely understand what's the real meanings of the numbers in each column. Some data handlings are also needed. When making a plot, the labels should be changed by full real name for people to read the figure better. Below is part of visualizations about respondents' demographics.

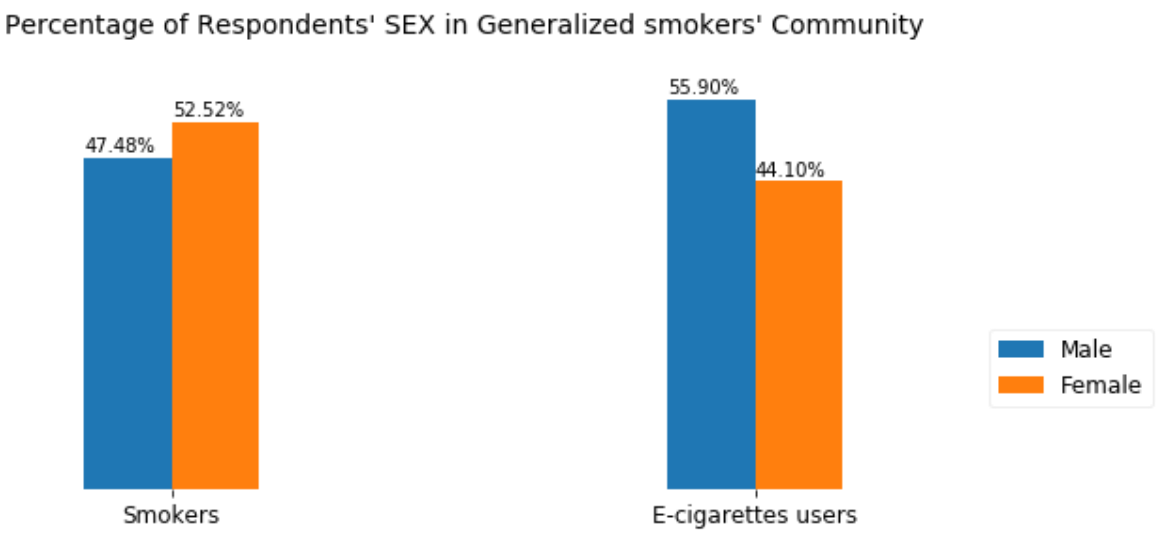

Fig. 1 Bar Chart of Respondent's Gender Distribution

The gender distributions of smokers and e-cigarettes users are quite different. The percentage of male e-cigarettes users are much higher than female users, while the percentage of male smokers is even lower than female smokers, which conflicts the public sense. Through the investigation, we found the number of female respondents from the original dataset is much more than male. This should be the reason leading to the unnormal outcome.

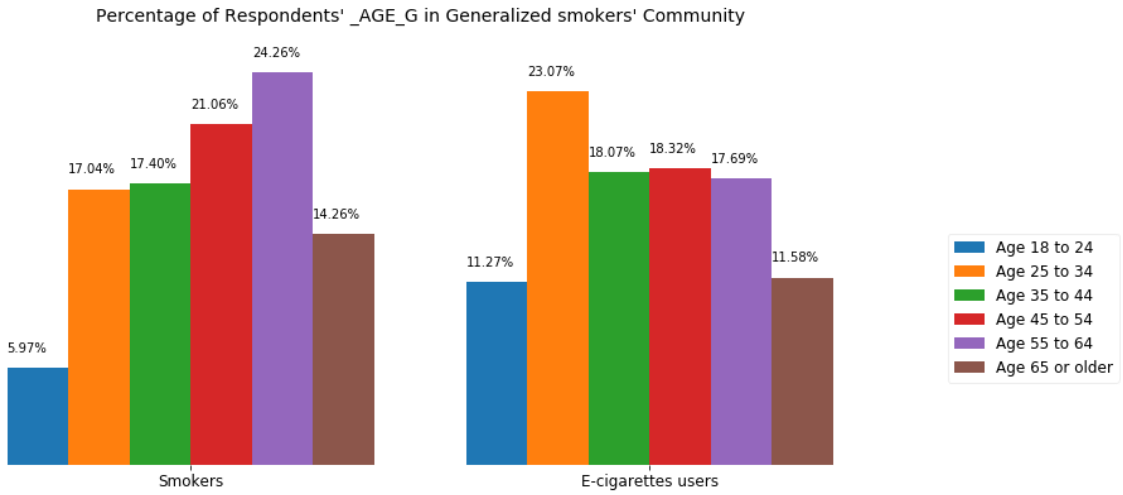

Fig. 2 Bar Chart of Respondent's Age Distribution

The distributions of age in both groups also have inconsistent characteristics. We can find middle-aged people (age ranges from 45 to 64) show less interest in using e-cigarettes. Younger people (age ranges from 18 to 34) show greater preference in using e-cigarettes. No obvious differences were detected from both groups at age intervals [35,44] and [65,].

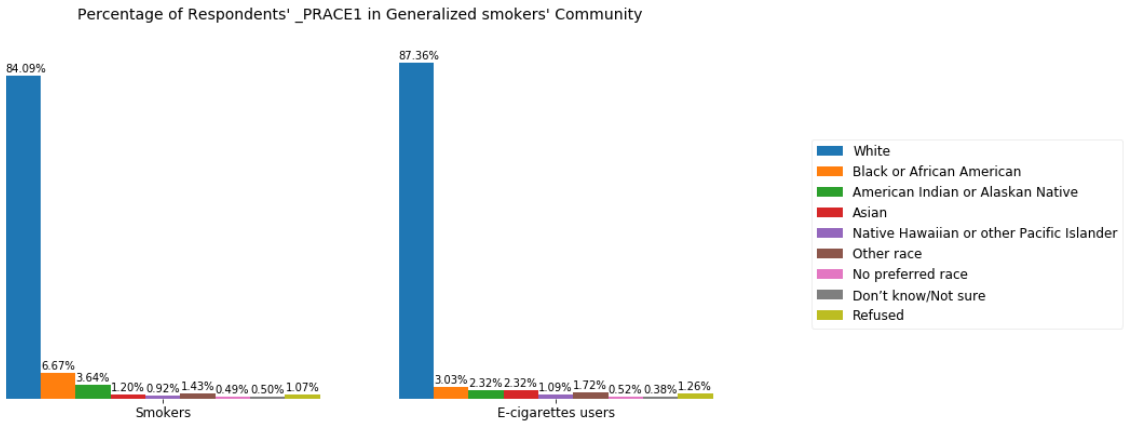

Fig. 3 Bar Chart of Respondent's Race Distribution

The distributions of race in both groups are very similar. Black or African American people show slight preference to traditional cigarettes comparing with its low percentage closes to other races' percentages. The percentage of white people in the original dataset is up to 80.84%. This prevent us from making too credible conclusions about white people's smoking preference based on this bar chart.

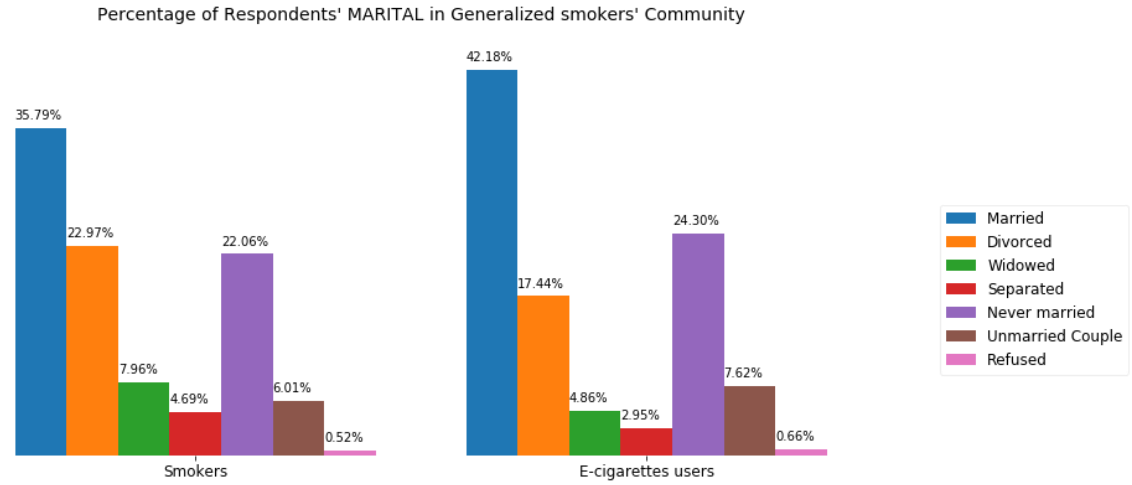

Fig. 4 Bar Chart of Respondent's Marital Distribution

The distribution of marital situation in both groups show a higher percentage of people who have incomplete marriage (divorced, widowed, separated) tend to smoke traditional cigarettes. A higher percentage of people who are in marriage or never married tend to use e-cigarettes.

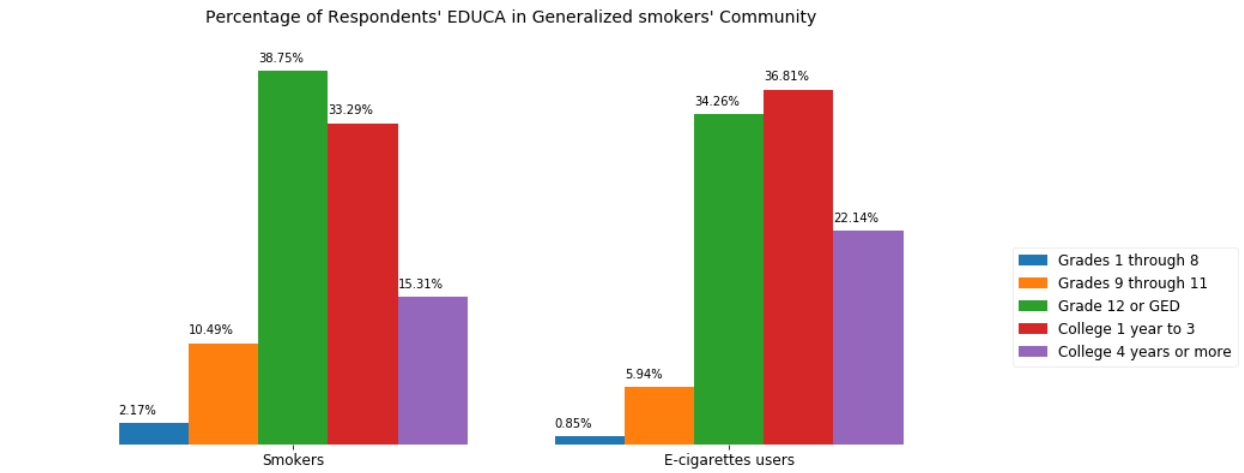

Fig. 5 Bar Chart of Respondent's Education Distribution

From the bar chart of respondents' education level, we noticed people with GED or lower education prefer smoking than electronic cigarettes. However, when the education rises to college or higher, the trend of electronic cigarettes has risen significantly. Also, it's noticed that people with an education level between grade 12 and college 4 year is the main force for both groups. We can surmise the reason why percentage goes down in college 4 years or more is because many respondents haven't reached to that level of study rather than they didn't.

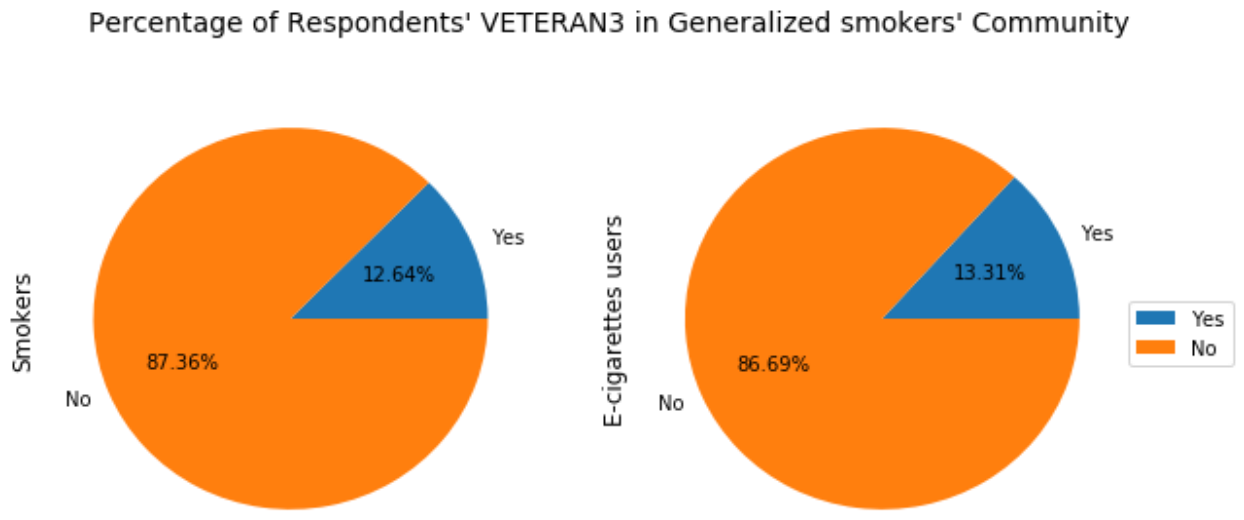

Fig. 6 Pie Chart of Respondent's Veteran Distribution

The above pie chart of percentage of veteran situation in both groups indicates there's no significant difference among smokers and e-cigarettes users.

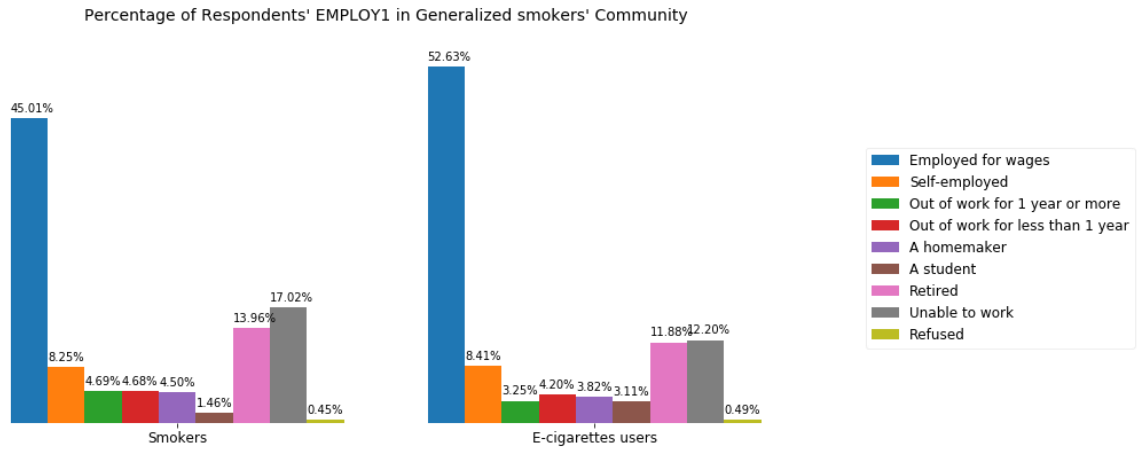

Fig. 7 Bar Chart of Respondent's Employment Distribution

From the bar chart of percentage of respondents' employment information, the percentage of people employed for wages and students in e-cigarettes users are higher than in smokers. The percentages of people who are retired or unable to work in smokers are higher than in e-cigarettes users.

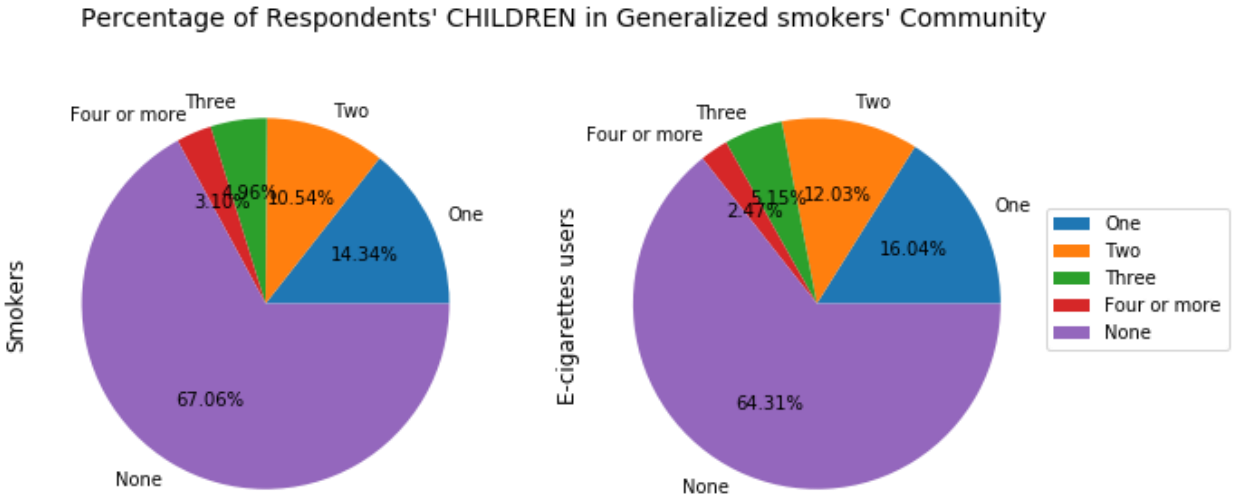

Fig. 8 Bar Chart of Respondent's Children # Distribution

From the above pie chart of percentage of respondents' children number in both groups, we didn't find clear differences among smokers and e-cigarettes users. One interesting finding is that the percentage of people who have at least one child in smokers is slightly higher than in e-cigarettes users, which means electronic smoking products have become common choices among people who have children.

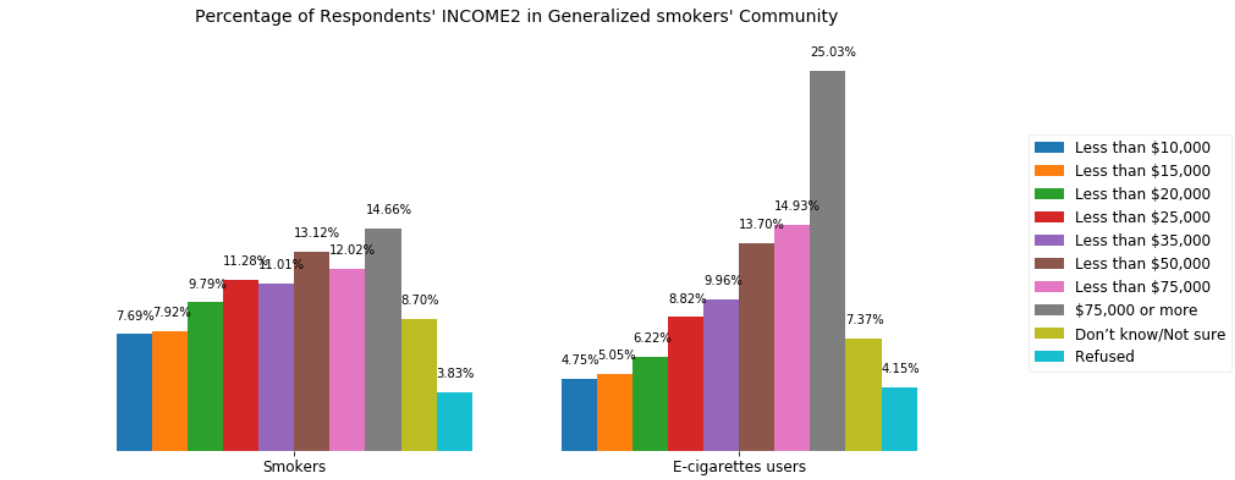

Fig. 9 Bar Chart of Respondent's Income Distribution

The bar chart of respondents' income distribution shows a strong contrast between smokers and e-cigarettes users. The trend of smoking in smokers increases gradually with the increase of people's income, which can properly interpret people with higher income also have greater tobacco purchasing power. However, this regularity looks still uncertain comparing with the trend of e-cigarettes smoking in e-cigarettes users. We discovered the percentage of people who use e-cigarettes follows a strict geometric growth trend. People with higher income have more desire to buy e-cigarettes and the rate of growth also becomes more dramatic when the level of income is higher. This leads us to think of if e-cigarettes have had some special values beyond the product itself, such as sense of identity, personal taste or spending power and so on.

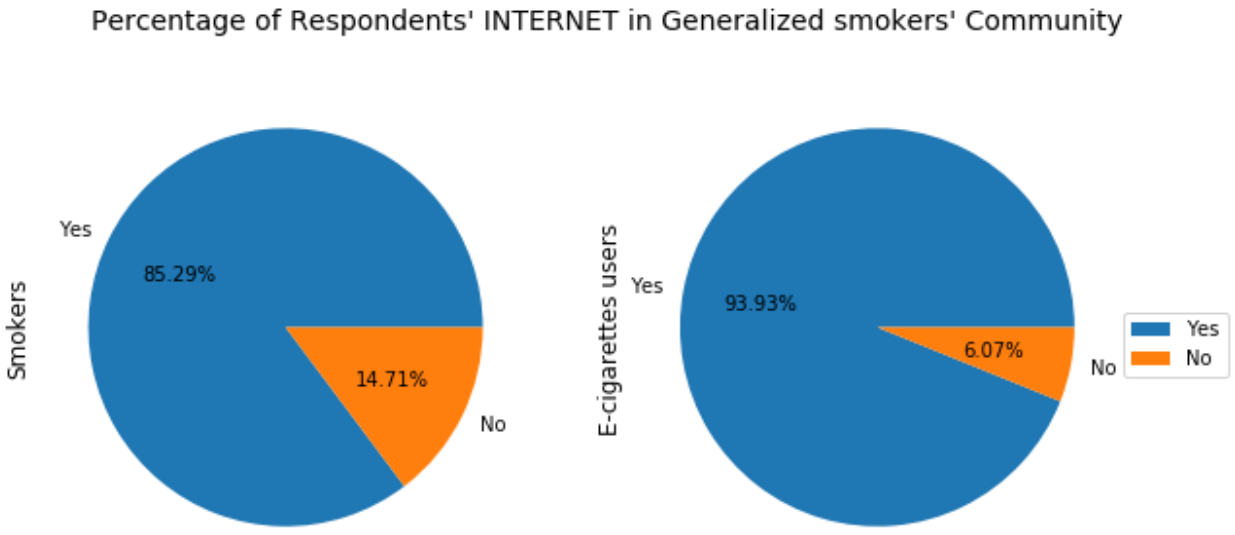

Fig. 10 Pie Chart of Respondent's Internet Use Distribution

This pie chart above shows that the main body of e-cigarettes users are commonly able to use Internet and use it very often. Comparing with it, there are still 14.71% of the smokers don’t use Internet as a daily tool.

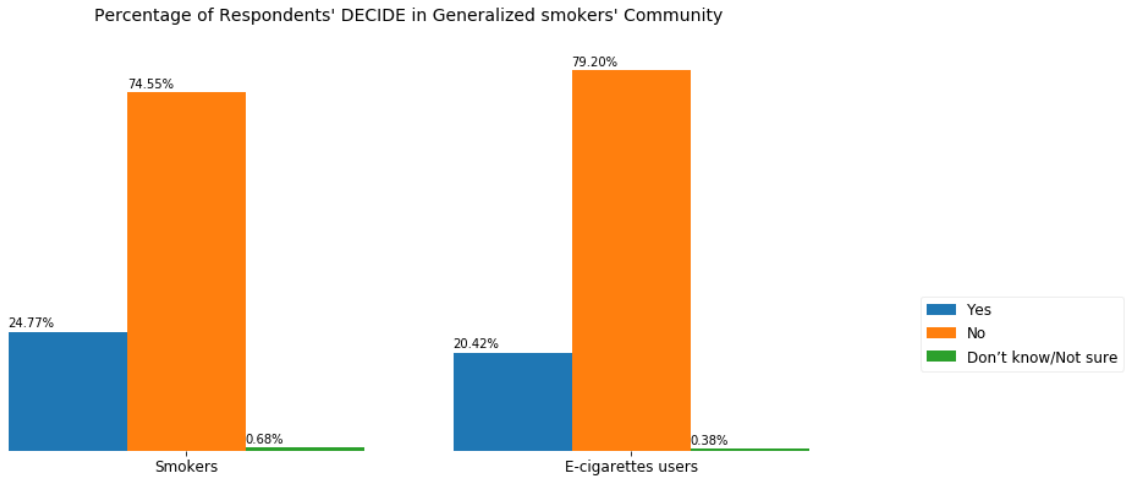

Fig. 11 Bar Chart of Respondent’s Cognitive Ability Distribution

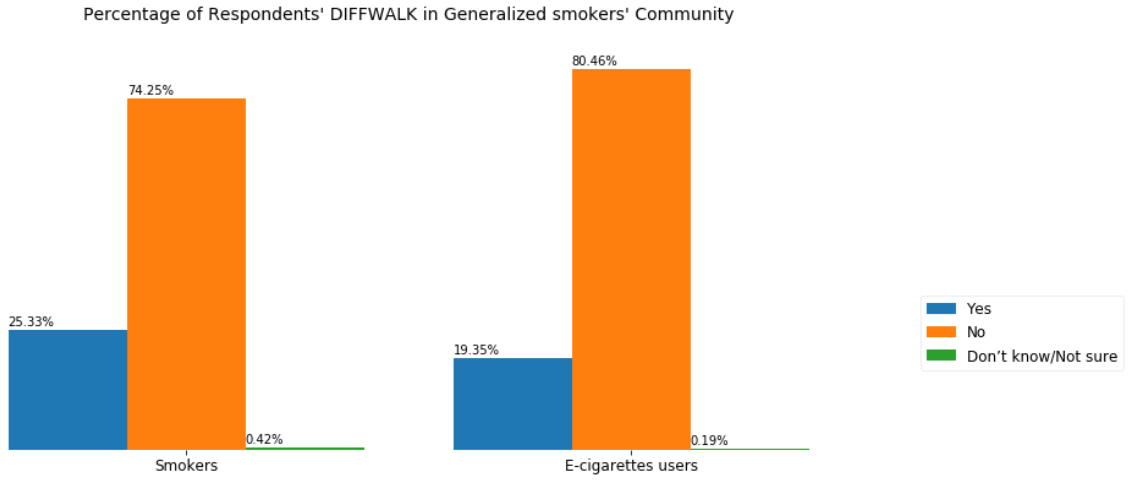

Fig. 12 Bar Chart of Respondent’s Moving Ability Distribution

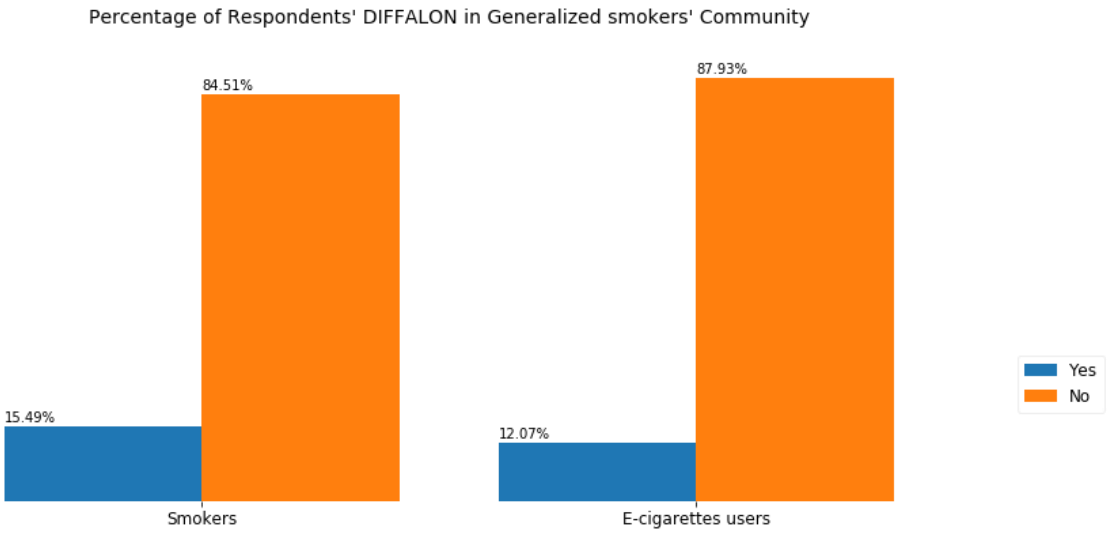

Fig. 13 Bar Chart of Respondent's Physical & Mental Health Distribution

The bar charts of respondents' cognitive ability, moving ability and physical and mental status above shows people use e-cigarettes are less likely having a cognitive ability disorder, moving difficulty, physical disorder or mental disorder, which can prove from the side that e-cigarettes have less damage to human.

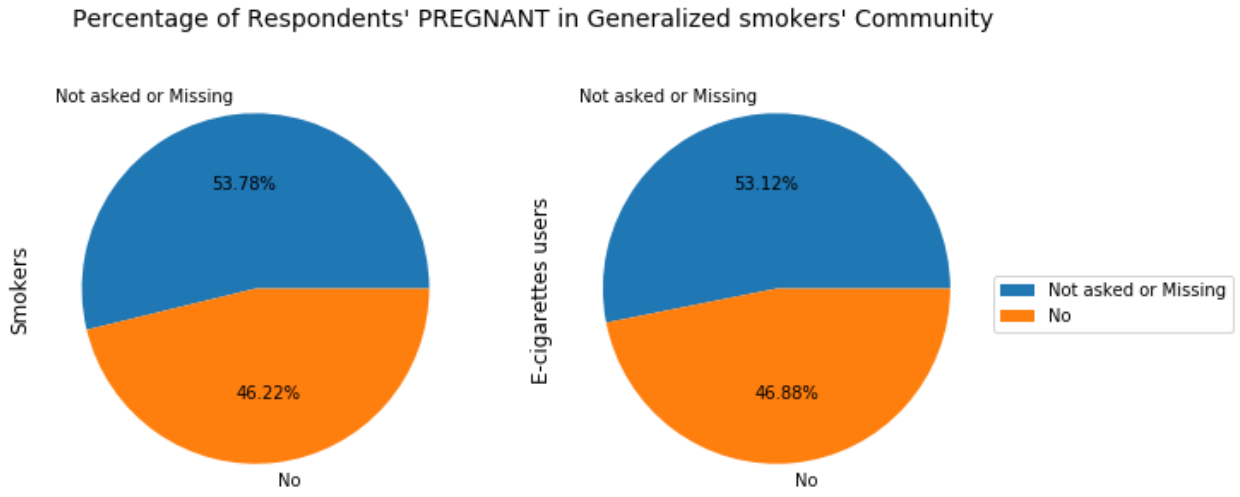

Fig. 14 Pie Chart of Respondent's Pregnancy Distribution

Finally, we compare the percentages of pregnant people in both groups of respondents. Only female respondents are counted in this analysis. The outcome shows the pregnancy isn't affected by smoking preference significantly.

Ⅲ Statistics Tests for Binary-Liked Variables

In this section, we conducted Chi-Squared tests on many variables we retrieved from the original dataset. These variables share a common characteristic, their answer options are the same or similar. In these variables, almost every number showen as 1 represent a “Yes” while every number shown as 2 represent a “No”.

Table. 4 P-values of Variables from Chi-Squared Tests

	P_value	Adjusted_p_value	Rejection
COPDCOGH	5.624244e-29	4.611880e-27	True
COPDFLEM	8.251870e-25	3.383267e-23	True
CHCCOPD1	2.737552e-24	7.482641e-23	True
COPDBRTH	7.186029e-13	1.473136e-11	True
ASTHNOW	5.212181e-10	8.547977e-09	True
HAVARTH3	6.893594e-10	9.421246e-09	True
HLTHPLN1	5.515026e-08	6.460459e-07	True
CVDINFR4	9.733486e-08	9.976823e-07	True
DIABETE3	8.247240e-06	7.514152e-05	True
MEDCOST	2.067987e-05	1.695749e-04	True
TOLDHI2	1.860664e-04	1.387040e-03	True
SHINGLE2	2.564010e-04	1.752073e-03	True
COPDBTST	4.235136e-04	2.671393e-03	True
CVDCRHD4	4.931786e-04	2.888617e-03	True
ARTHWGT	8.165699e-04	4.463915e-03	True
FLUSHOT6	1.613728e-03	7.854180e-03	True
CHCOCNCR	1.628306e-03	7.854180e-03	True
CVDSTRK3	3.384284e-03	1.541729e-02	True
ARTHEDU	4.671185e-03	2.015985e-02	True
SDHBILLS	7.082893e-03	2.903986e-02	True
CVDASPRN	8.902831e-03	3.476343e-02	True

About 20 variables are filtered out that their values are significantly related with people’s choice of smoking. These finding are extremely meaningful to support further deeper study on the correlation of a single variable and smoking preference. From the table. 4, we noticed the top five variables are all COPD related variables or asthma related variables. Other findings from the statistical tests also include the different impact on arthritis, rheumatoid arthritis, gout, lupus, fibromyalgia, heart attack (as called a myocardial infarction), diabetes, high blood cholesterol,

angina or coronary heart disease, cancer (any type) and stroke between smokers and e-cigarettes users.

Besides, people with different choices of cigarettes products also behaves differently, such as their actions on having a health coverage, paying for medical cost, taking a shingles or zoster vaccine, losing weight for arthritis or joint symptoms, taking a flu shot, taking educational course or class for arthritis or joint symptoms, not being able to pay mortgage, rent or utility bills, taking aspirin, etc.