**The Correlation between Youth Tobacco Consumption and Mortality Rate**

**Course:** Business Analytics for Manager

**Group:** Pyramid

**Professor:** Prof. RP

**Team Members:**

Bingyue Zeng

Jie Lu

Qilin Liu

Xin Sui

**Date:**

April 27th, 2020

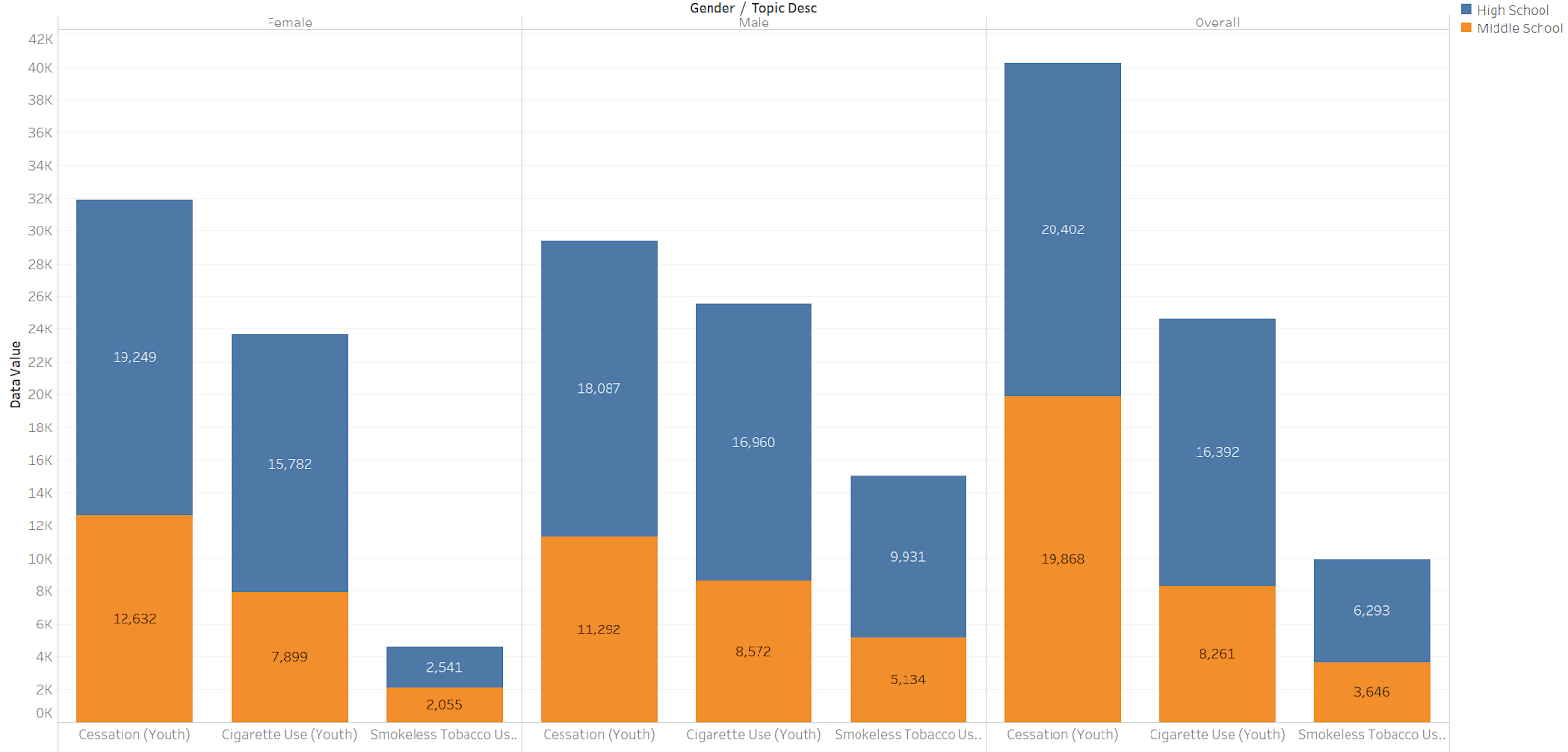
1. **Matrixes of Storytelling**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis** | **Question** | **Analytic type** | **Variable** | **Chart** |
| Smoking rate is associated with different demographic segmentations such as gender, education and smoking type | Do different gender, education level, and smoking type have the same smoking rate? | Descriptive | Gender, Education, TopicDesc, Smoke\_Rate | Stacked Bar Chart |
|  | Does the observation in gender balance in each state among different smoking types? | Descriptive | Gender, Education,  STATE, TopicDesc, Smoke\_Rate | Grouped Bar Chart |
|  | Do the 3 classes in smoking type balance? | Descriptive | TopicDesc | Pie Chart |
|  | What is the trend of common cigarette and smokeless cigarette in each state? | Descriptive | State, cigarette rate, smokeless rate | Treemap |
|  | In each education level, does the smoking rate vary too much on different smoking types? | Descriptive | Gender, Education,  TopicDesc, STATE, Smoke\_Rate | Box-and-Whisker Plot |
| The distribution of smoking rates follows a normal distribution. | Is the standard deviation of the smoking rate Gaussian distributed? | Descriptive | Smoke\_Rate | Histogram |
|  | Is there any difference between the middle school’s smoking rate distribution and the high school’s smoking rate distribution? | Descriptive | Education, Smoke\_Rate | Faceted Histogram |
| The mortality rates (cancer, heart disease, drug overdose, stroke) are correlated to each other. | Is there any correlation among various mortality cause? | Descriptive | CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Scatter Plot |
|  | Is there an association between death counts and sample sizes? | Descriptive | CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | line chart |
|  | Is there any mortality rate record unusually high/low? | Descriptive | CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Box and Whisker Plots |
| The 4 types of mortality rate (cancer, heart disease, overdose and stroke) vary through time | Which smoke type has the highest mortality rate from 2014 to 2017? | Descriptive | YEAR,CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Horizontal Bar Chart |
|  | What will 4 types of mortality  Rate (cancer, heart disease, overdose and stroke) be like in the future? | Predictive | YEAR,CANCER\_DEATHS, HEART\_DISEASE\_DEATHS, OVERDOSE\_DEATHS, SREOKE\_DEATHS | Line Chart |
|  | What will happen to the mortality rate of 4 different diseases in NJ after 2017? | Predictive | YEAR, STATE,  HEART\_DISEASE\_RATE,  OVERDOSE\_RATE, STROKE\_RATE, CANCER\_RATE | Continuous Line Chart and Trend Line |
|  | What will happen to the mortality rate of 4 different diseases in NE after 2017? | Predictive | YEAR, STATE,  HEART\_DISEASE\_RATE,  OVERDOSE\_RATE, STROKE\_RATE, CANCER\_RATE | Continuous Line Chart and Trend Line |
| The mortality rate of heart disease varies among youth with different gender and education levels. | What is the mortality rate of 4 diseases among youth with different gender? | Descriptive | YEAR,CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Horizontal Bar Chart |
|  | What is the mortality rate of heart disease among youth with different gender and education levels? | Descriptive | YEAR,HEART\_DISEASE\_RATE, Education,Gender | Stacked Bubble Chart |
|  | What is the mortality rate of heart disease and cancer among youth with different education levels? | Descriptive | YEAR,CANCER\_RATE, HEART\_DISEASE\_RATE,Education | Gantt Chart |
| Smoke rate is positively correlated with mortality rate | What is the distribution of smoke rate and the proportion of death among population in each state? | Descriptive | Latitude,  Longitude,  State, death\_prop, smoking status death rate | Heat Map |
|  | For each state, what is the trend of smoke rate and total mortality rates from 2014 to 2017? | Descriptive | Year, State, Smoke\_Rate, CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Scatterplot |
| The willingness of cessation is associated with higher rate of smoking | What is the association between cessation and smoke rate? | Descriptive | State, smoke\_rate, cessation smoke rate, non-cessation smoke rate | Horizontal Bar Chart |

One Statistical Model & One Machine Learning Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Question** | **Analytic type** | **Variable** | **Model** |
| Statistical Analysis | Is there any significant difference among the smoking rates for different smoking types? | Descriptive | TopicDesc,CANCER\_RATE, HEART\_DISEASE\_RATE, OVERDOSE\_RATE, SREOKE\_RATE | Model: ANOVA test |
| Machine Learning Analysis | Given all independent variables, can we predict whether a states’ smokers have a high or low cancer mortality rate. | Predictive | YEAR，STATE，TopicDes， MeasureDesc， Smoke\_Rate， Data\_Value\_Std\_Err, Sample\_Size, Gender, Education, HEART\_DISEASE\_RATE, HEART\_DISEASE\_DEATHS, OVERDOSE\_RATE, OVERDOSE\_DEATH, SREOKE\_RATE, STROKE\_DEATHS, CANCER\_RATE, CANCER\_DEATHS | Model: Logistic Regression |

1. **Analytics Charts (following matrix row sequence)**
2. Bar Chart for Youth Smoking Frequency - Jie Lu

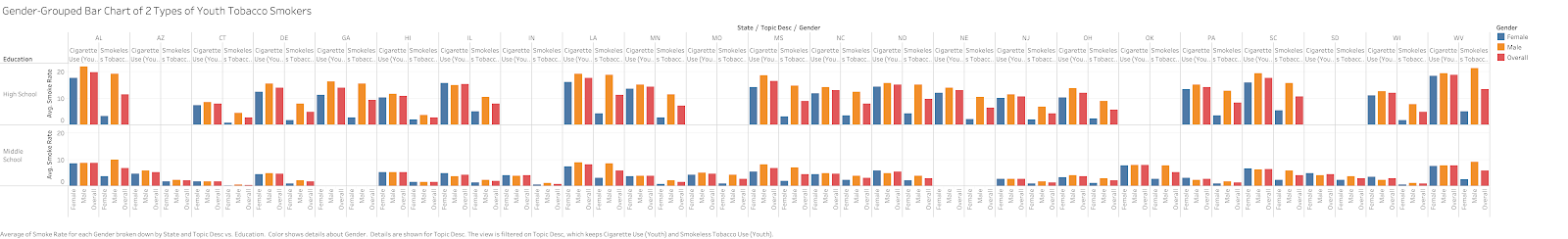
Figure1: Bar chart for Youth Smoking Frequency

**A**: This figure is a bar chart showing the frequency of youth smoking status, segmented by gender, education and smoking type.

**B**: The height of the bar represents frequency, each facet represents gender (female, male, overall), each bar within each facet column indicates smoking type (cessation, tobacco, smokeless) and color differs between high school and middle school.

**C**: It shows that youth male smokes more than female, high school students smokes slightly more than middle school students and most students who ever smoked are in the cessation program.

1. Gender-Grouped Bar Chart of 2 Types of Youth Tobacco Smokers in 23 States - Bingyue Zeng



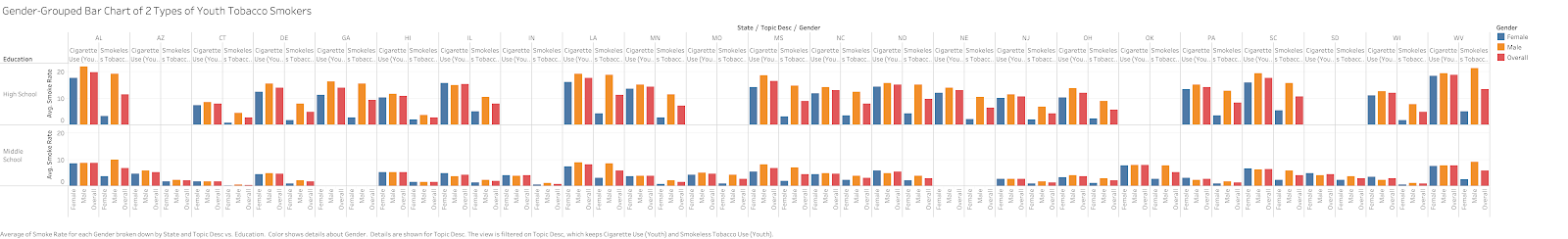


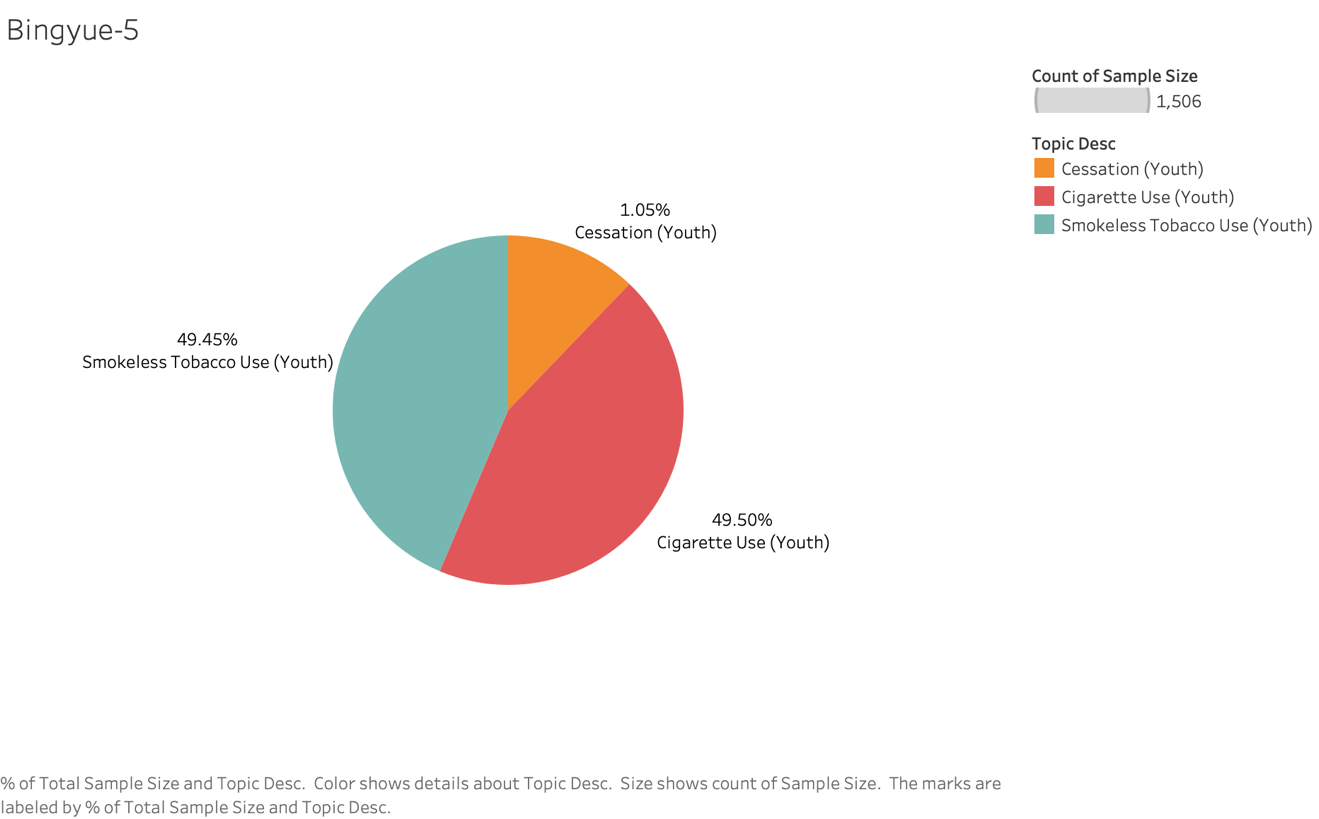
Figure2: Gender-Grouped Bar Chart of 2 Types of Youth Tobacco Smokers in 23 States

**A:** This figure shows a gender-grouped bar chart that depicts the smoking rate distribution among gender, smoking type, and education level in 23 states.

**B:** This figure drills down to the education level that the smoking severity of both cigarette and smokeless cigarettes in high school is much higher than that in middle school in the United States. It also drills down to the gender level that male smokers usually have higher smoking severity over that of female smokers in high school across the country, and male and female smokers are of similar smoking severity in middle school across the country.

**C:** It implies that state government and educational institutions should put more emphasis on high school male smokers when enforcing tobacco restriction.

1. Pie Chart of 3 Classes in Smoking Type - Bingyue Zeng

Figure3: Pie Chart of 3 Classes in Smoking Type

**A**: This figure is a pie chart showing the percentage of the observations for each smoking type.

**B**: The ‘Smokeless Tobacco Use’ proportion and ‘Cigarette Use’ are similar, around 49%, but ‘Cessation’ is not.

**C**: It implies that there imbalance classes exist among 3 types of tobacco use. The study focus on just smokeless tobacco and cigarette use should be better.

1. Treemap of the average smokeless user smoke rate and average cigarette smoke rate in recorded states during 2014-2017 - Qilin Liu

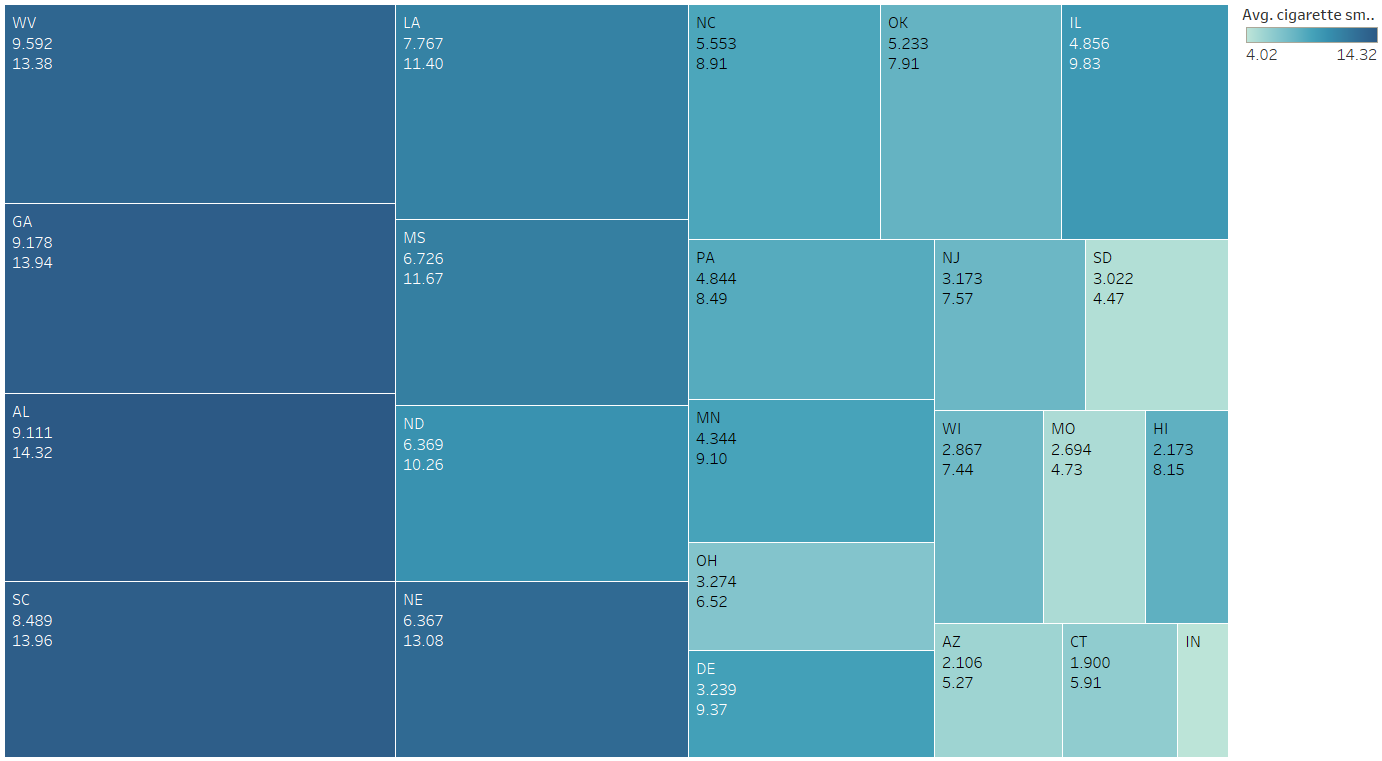


Figure4: Treemap of the average smokeless user smoke rate and average cigarette smoke rate in recorded states during 2014-2017

A. This figure is a treemap. The larger the cube means higher average smokeless user average smoke rate, and deeper color refers to higher average common cigarette smoke rate.

B. The states with high average smokeless smoke rate also have a relatively high average cigarette smoke rate.

C. The government should notice that smokeless tobacco will not decrease the usage of common cigarettes, on the contrary, they might be positively associated.

1. Box-and-Whisker Plot of 2 Types of Youth Tobacco Smokers - Bingyue Zeng

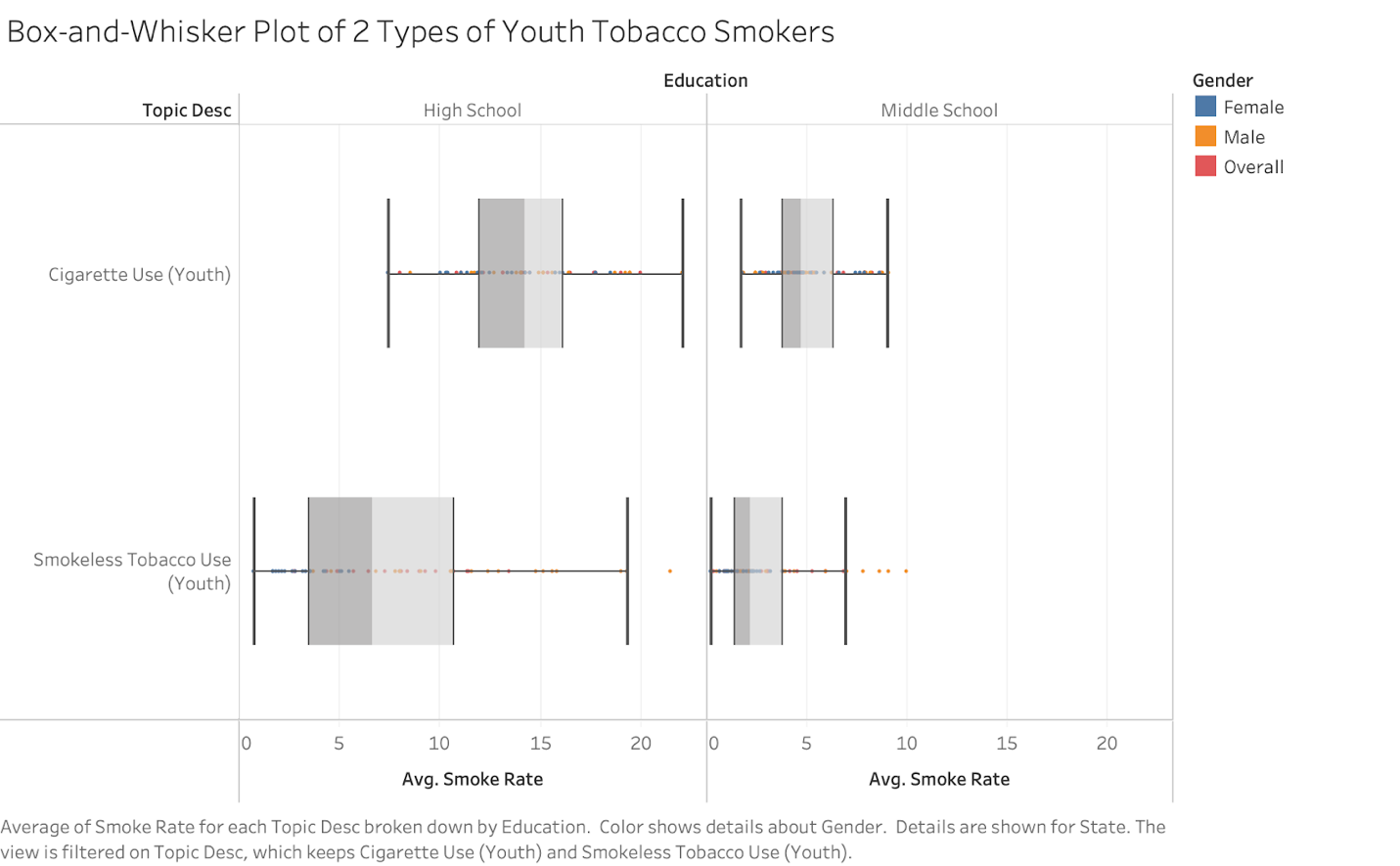


Figure5: Box-and-Whisker Plot of 2 Types of Youth Tobacco Smokers

**A:** This figure shows the box and whisker plots of the average smoking rate between 2 types of youth tobacco smokers in different education levels.

**B:** The average smoking rate varies in high school cigarette smokers, high school smokeless tobacco smokers, middle school cigarette smokers, and middle school smokeless tobacco smokers. For smokeless tobacco use, high school and middle school have some tiny proportion of the outliers above the quantile.

**C:** The findings indicated that a careful preprocessing process such as removing the outliers could be helpful to the analysis.

1. Histogram of the Overall Smoking Rate - Jie Lu

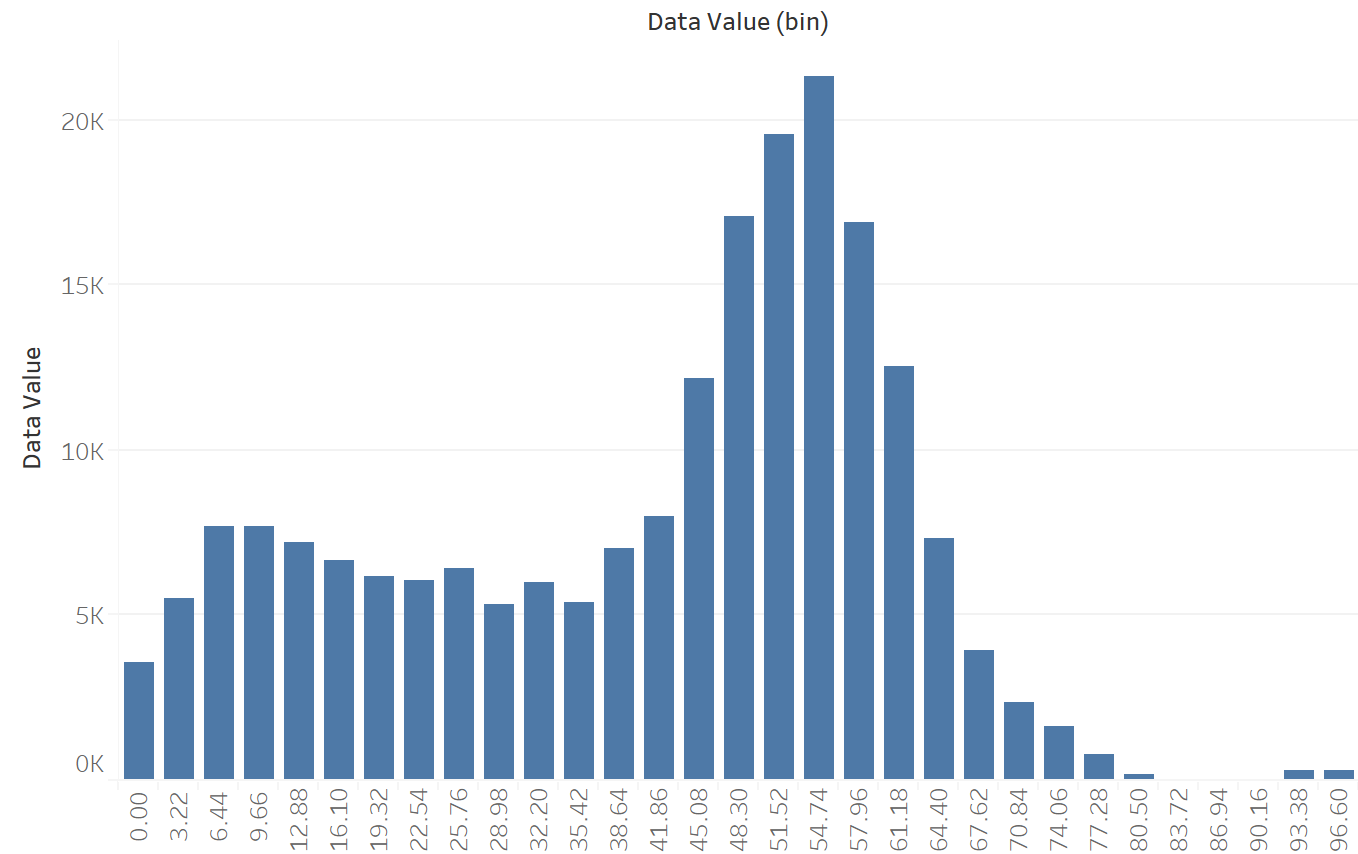
****

Figure6: Histogram of the Overall Smoking Rate - Jie Lu

**A:** This figure shows the histogram of the overall smoking rate in the survey.

**B:** The distribution is skewed towards the right side, and therefore didn’t follow a Gaussian distribution pattern.

**C:** This slight imbalance in the dependent variable is trivial, but indicated that the aggregate behavior of smoking --- most people smoked on an irregular basis, while a small portion of people smoked quite often.

1. Faceted Histogram of the Smoking Rate Across Education - Jie Lu

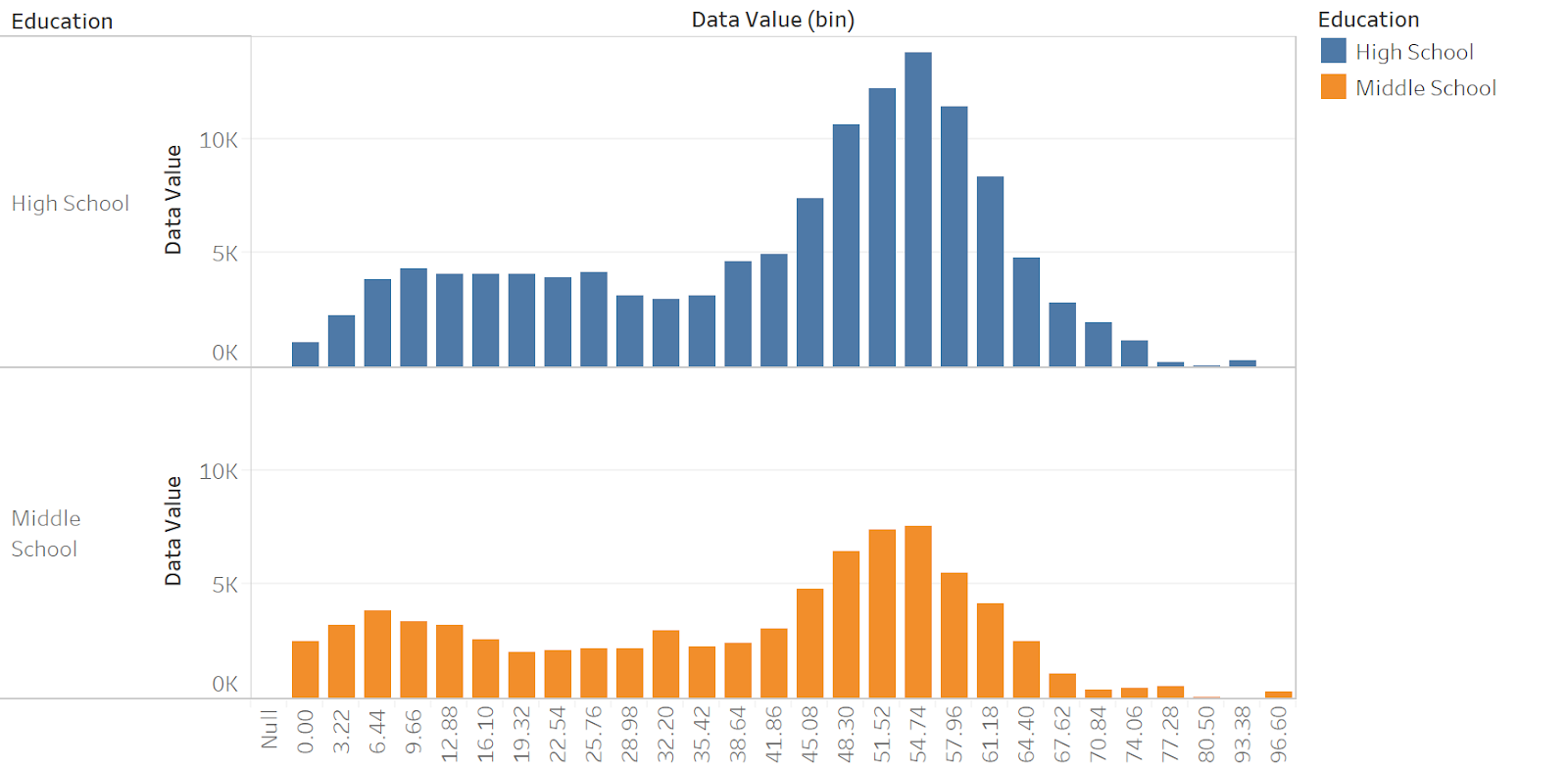


Figure7: Histogram of the Smoking Rate Across Education

**A:** This figure shows the faceted histogram of the smoking rate between high school and middle school.

**B:** Both high school and middle school shared a very similar distribution pattern. High school has a higher smoking rate among the whole range.

**C:** This implication means the youth’s smoking behavior didn’t change much across the education. However, there are much more smokers in high school.

1. Scatter plots of various mortality causes - Jie Lu

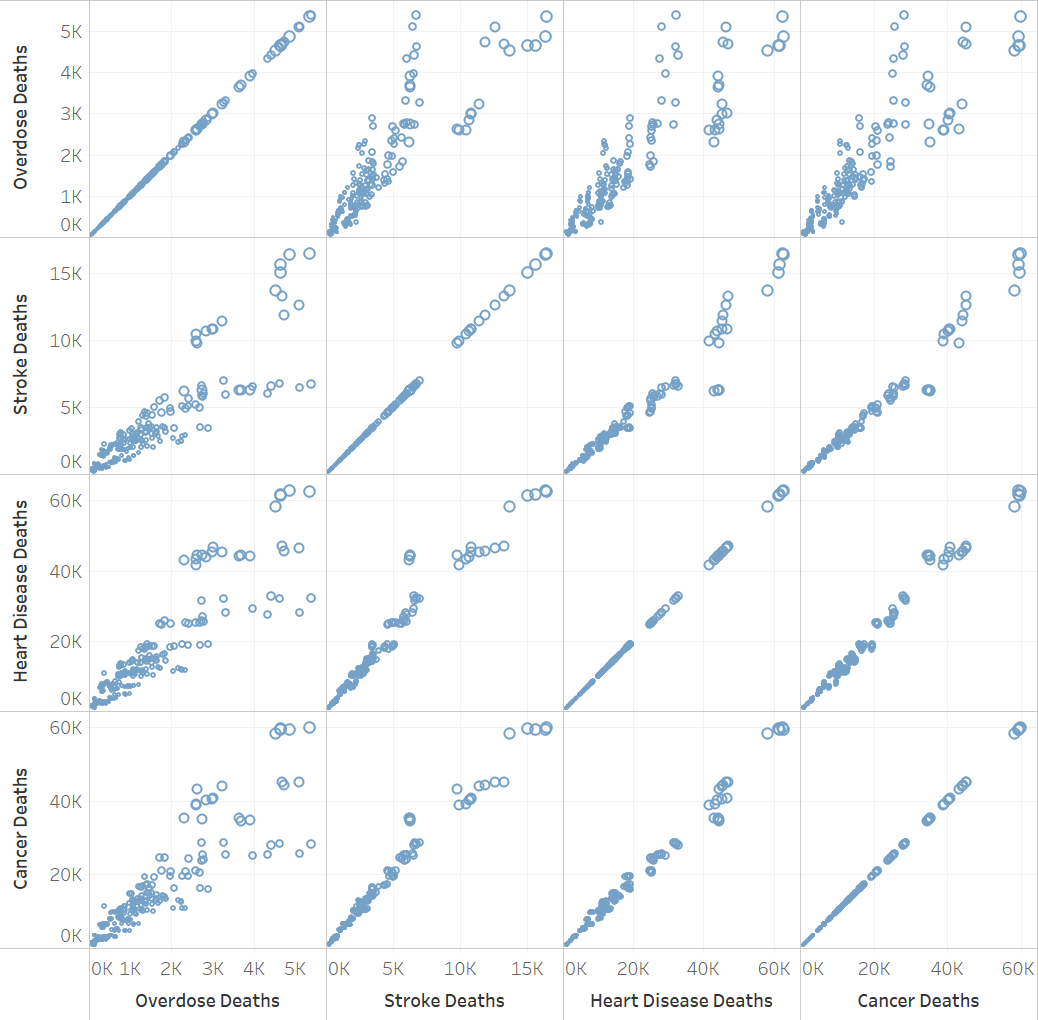


Figure8: Scatter plots of various mortality causes

**A**: This figure is a scatterplot matrix showing the correlation among various mortality causes.

**B**: The x-axis and y-axis include the same features, and each cell displays the scatter plot between two, e.g. cell(0, 1) shows the scatter plot between overdose and stroke.

**C**: Interestingly, those death causes are all positively correlated with each other.

1. Line Chart for the Sum of Samples and Death Counts in Smoke-related Mortality - Qilin Liu

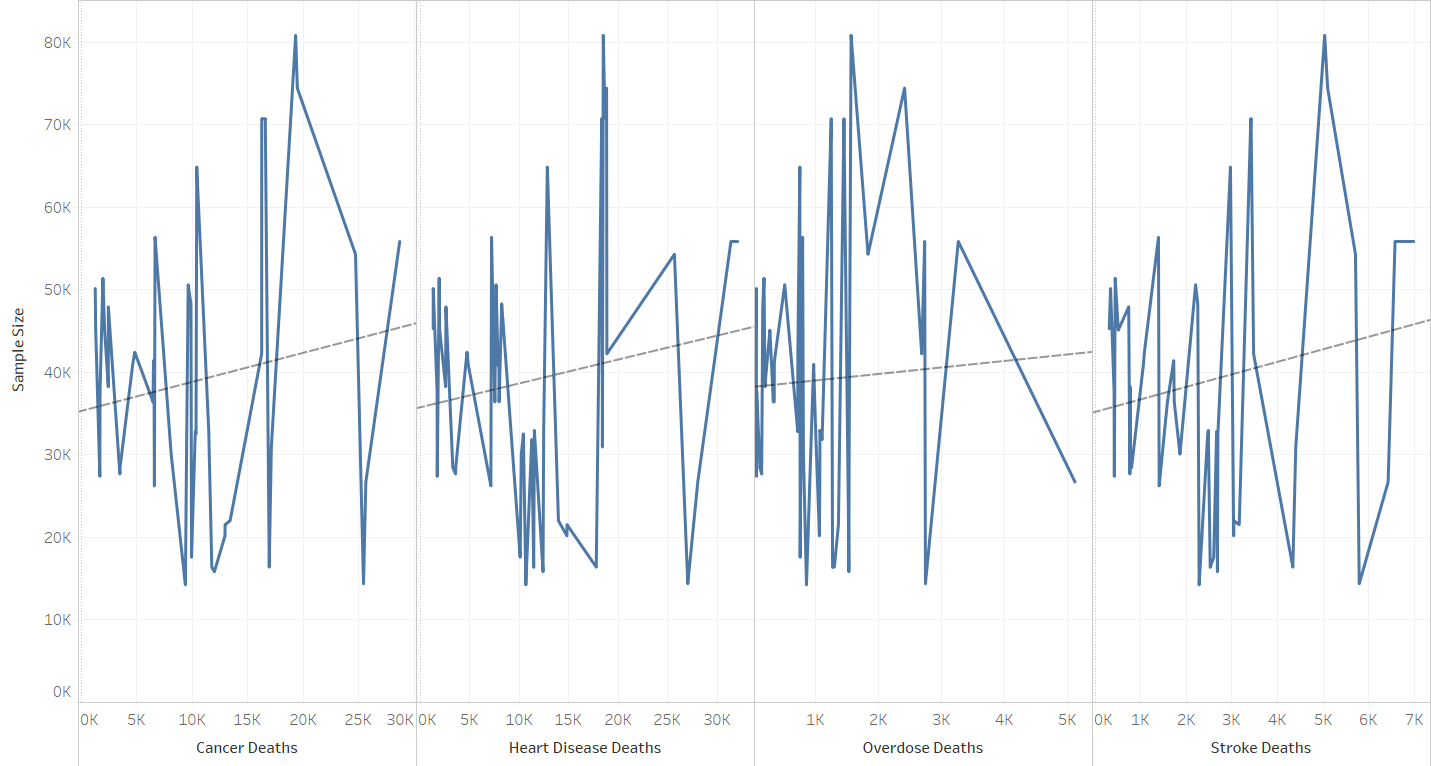


Figure9: line chart for the sum of samples and death counts in smoke-related mortality

A. This figure is a line chart. It shows there is no correlation between sample size and any smoking-related death counts. R-square and p-value are both too weak.

B. As all four types of smoking-related mortality have no correlation with the sample size, the death count does not simply rise as the population rises.

C. The researchers should weigh the local population for further death rate research.

1. Box Plots of Various Mortality Causes - Jie Lu

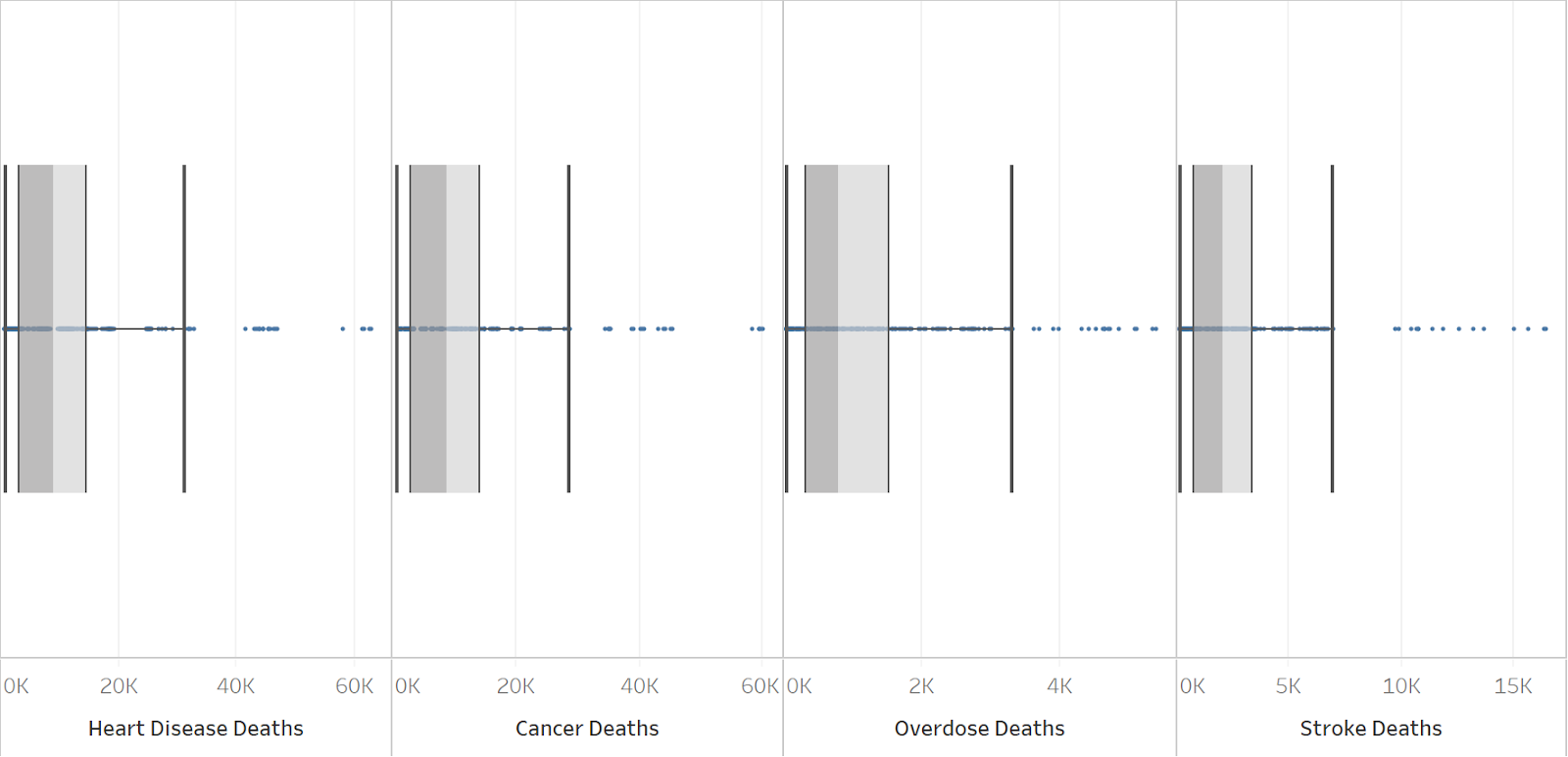


Figure10: Box plots of various mortality causes

**A:** This figure shows the box and whisker plots of the mortality causes.

**B:** Those mortality causes shared a very similar distribution, and all of them had a big portion of the outliers above the quantile.

**C:** The findings indicated that a careful preprocessing process such as removing the outliers could be helpful to the analysis.

1. Bar Chart of the Mortality Rate in Different Smoke Types – Xin Sui

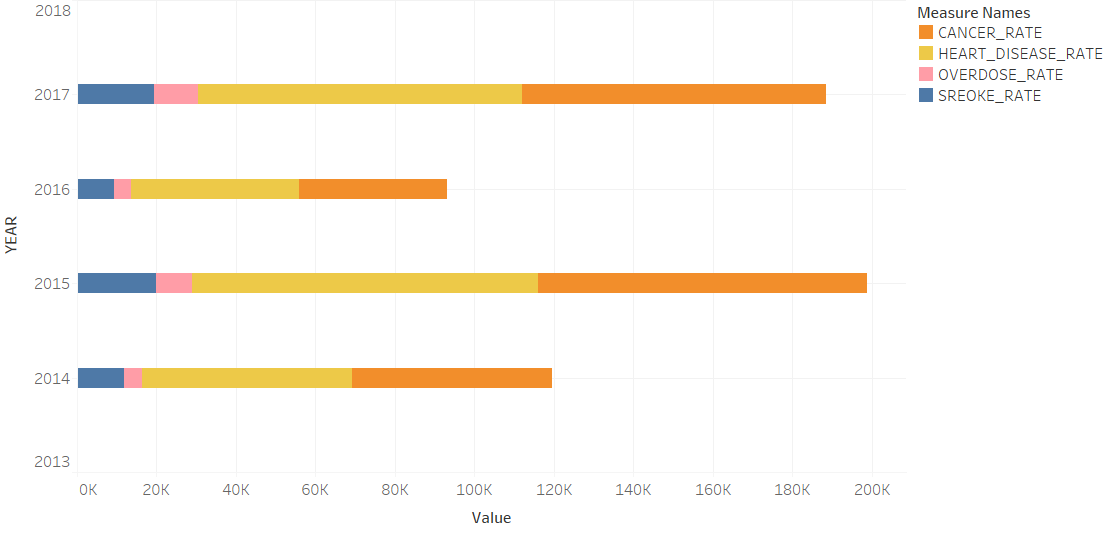


Figure11: Bar chart of the Mortality Rate in different smoke types

**A:** This figure above is the horizontal stacked bar chart showing mortality rates in different types of disease including cancer, heart disease, overdose and stroke, from 2014 to 2017.

**B:** In each year, the mortality rate of heart disease remains the highest, followed by that of cancer, stroke and overdose.

**C:** This implication means that people are under high risk of dying from heart disease, as well as under huge threat of cancer.

1. Line Chart of Different Mortality Causes for Each Year – Xin Sui

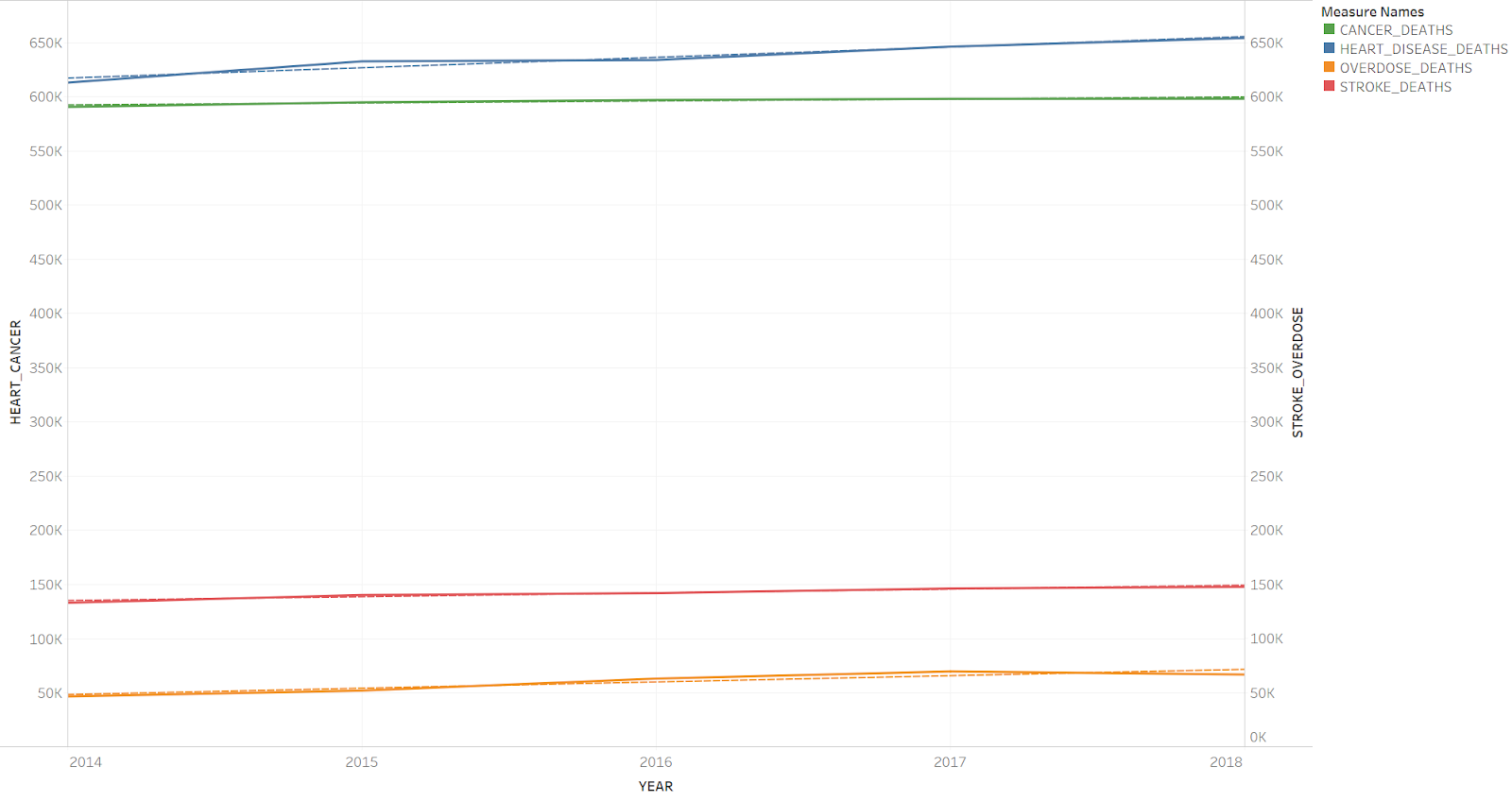


Figure12: Line chart of different mortality causes for each year

**A**: The line chart above is a time series chart connecting individual death cases in a view.

**B**: The number of death cases of heart disease ranked top, followed by cancer, stroke and overdose.

**C**: Overall, the number of cancer and that of heart disease are on a similar scale, far beyond stroke and overdose, and they all show a trend of steady growth over time.

1. Trend Line of 4 Types of Disease Rate in State NJ – Bingyue Zeng

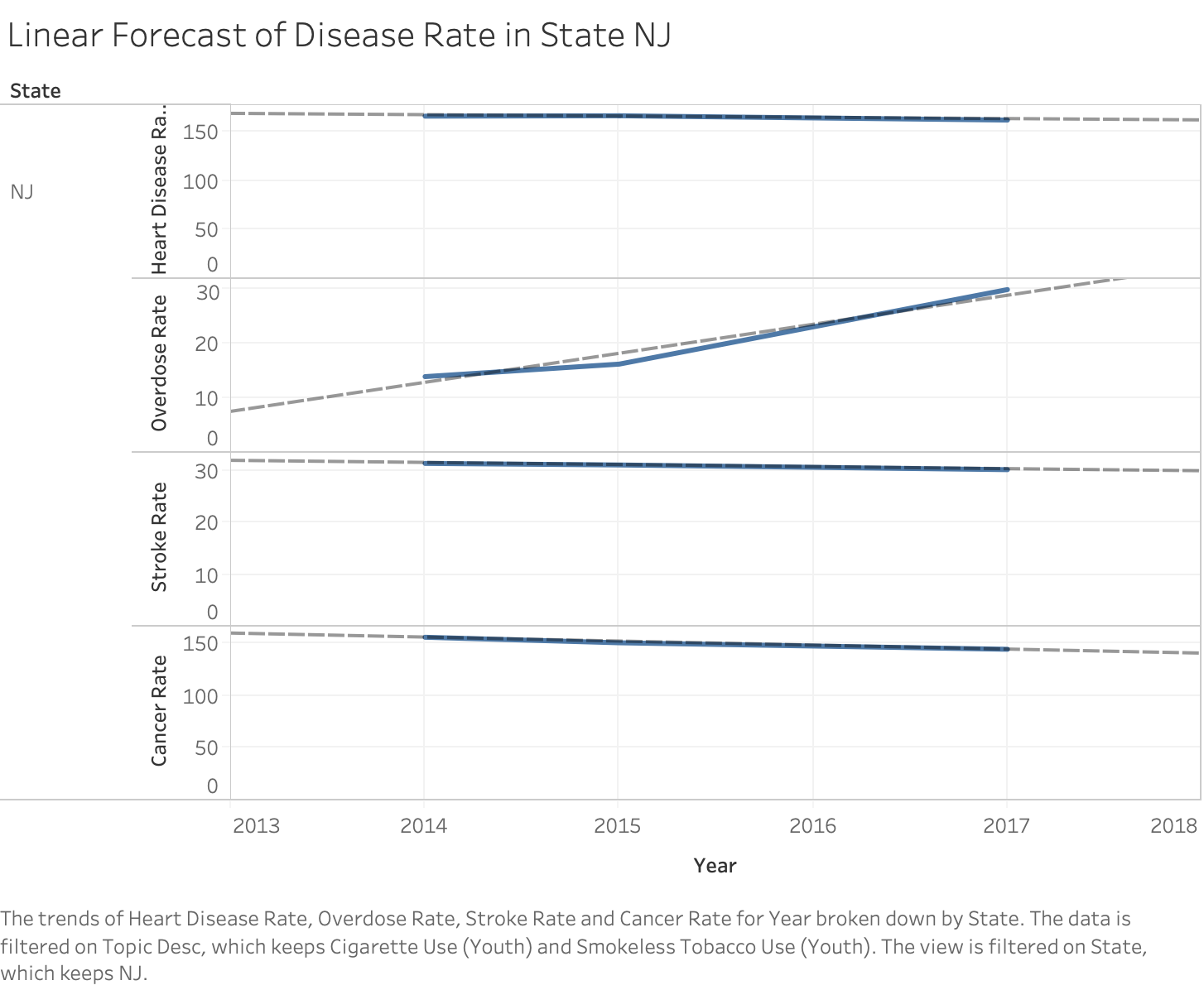


Figure13: Trend Line of 4 types of Disease Rate in State NJ

**A:** This figure shows trend lines that predict the death rate of heart disease, cancer, stroke and overdose in state NJ after 2017. R-squared are all above 74%.

**B:** Consumption in tobacco has the strongest positive correlation with drug overdose death rate. While heart disease death rate, cancer death rate, and stroke death rate each does not have positive correlation with tobacco smoking.

**C:** Smoking reliability in humans may increase drug addiction, such as opioids use disorder. Therefore, smoking a lot of tobacco is very likely to lead to an overdose death.

1. Trend Line of 4 Types of Disease Rate in State NE – Bingyue Zeng

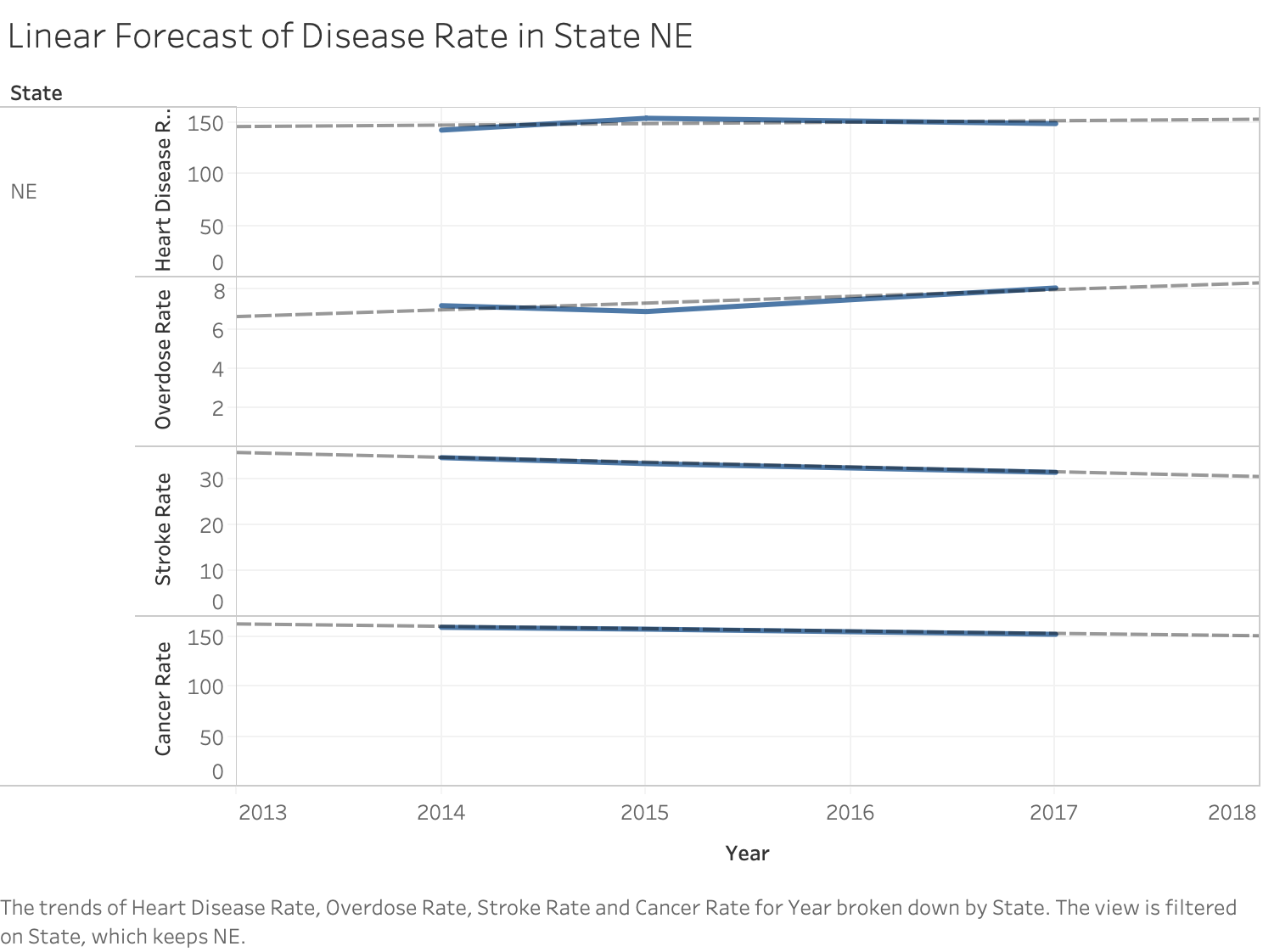


Figure14: Trend Line of 4 types of Disease Rate in State NE

**A:** This figure shows trend lines that predict the death rate of heart disease, cancer, stroke and overdose in state NJ after 2017. R-squared for overdose rate, stroke rate, and cancer rate are all above 70%

**B:** Consumption in tobacco has the strongest positive correlation with drug overdose death rate. While heart disease death rate, cancer death rate, and stroke death rate each does not have positive correlation with tobacco smoking.

**C:** Smoking reliability in humans may increase drug addiction, such as opioids use disorder. Therefore, Restriction on tobacco consumption is needed or more youth will die by drug overdose.

1. Stacked Bar Chart of Total Mortality Rates on Gender– Xin Sui

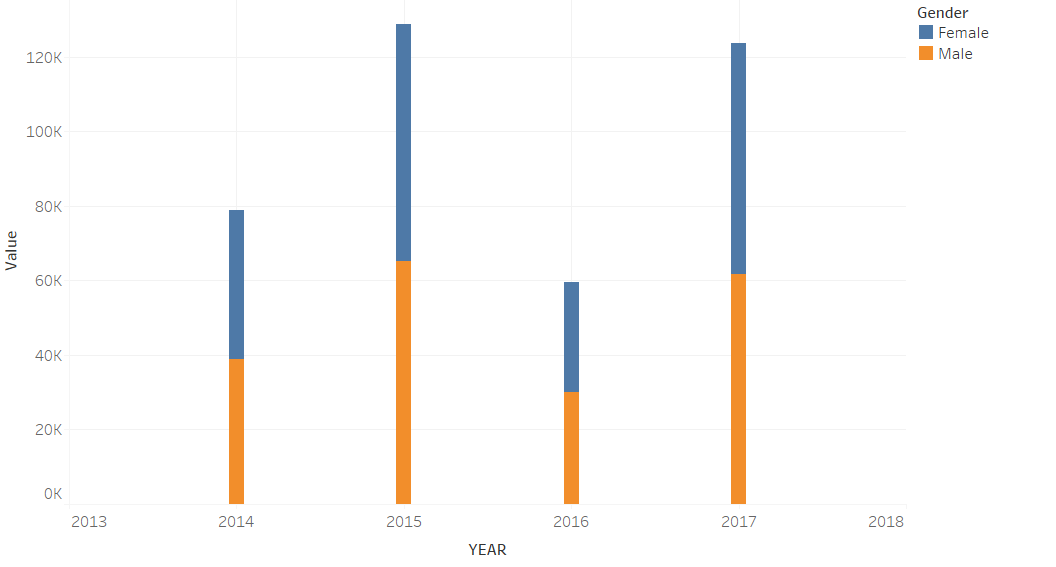


Figure15: Stacked Bar Chart of total mortality rates on gender

**A**: The bar chart above shows the total mortality rates on two genders in a view.

**B**: The total mortality rates of females and males are on a similar scale.  
**C**: It indicates that the risk of dying from tobacco consumption among the youth does not have a significant difference regarding gender.

1. Packed Bubble Chart of  Heart Disease Mortality Rate on Gender and Education– Xin Sui

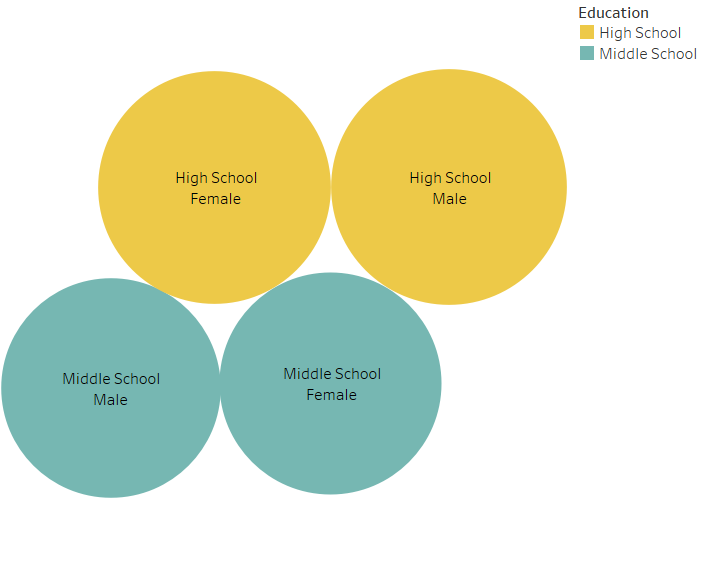


Figure16: Packed Bubble Chart of heart disease mortality rate on gender and education

**A**: The packed bubble chart above shows the heart disease mortality rate on two genders and education levels in a view.

**B**: The mortality rates of heart disease are quite similar in the lense of gender while high school youth shows a higher rate of death.

**C**: It indicates that the risk of dying from heart disease is higher among high school youth. It might somehow contribute to the fact that teenagers will develop tobacco dependency and die early if they are exposed to tobacco at an early age.

1. Gantt Chart of Mortality Rate of Two Disease on Education– Xin Sui

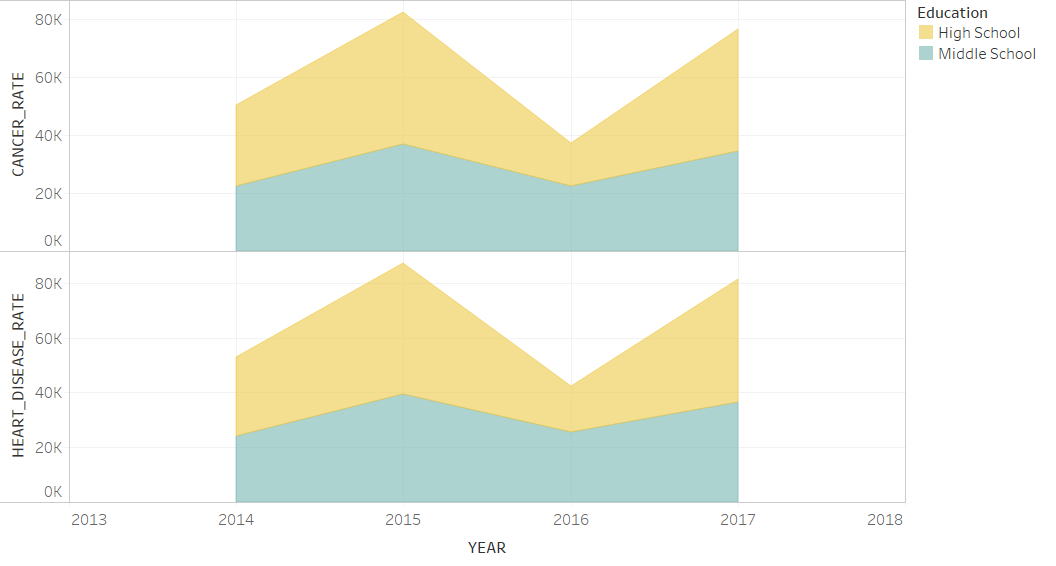


Figure17: Gantt Chart of mortality rate of two disease on education

**A**: The Gantt chart above shows the mortality rate of heart disease and cancer on two education levels in a view.

**B**: The mortality rates of heart disease and cancer are overall higher among high school youth than among middle school youth. In 2016, there was a steady drop-down in both mortality rates of high school students.

**C**: It indicates that high school students are under higher risk of dying from the two most dangerous diseases, which are heart disease and cancer.

1. Average Death Rate/Smoke Rate Distribution Heat Map - Qilin Liu

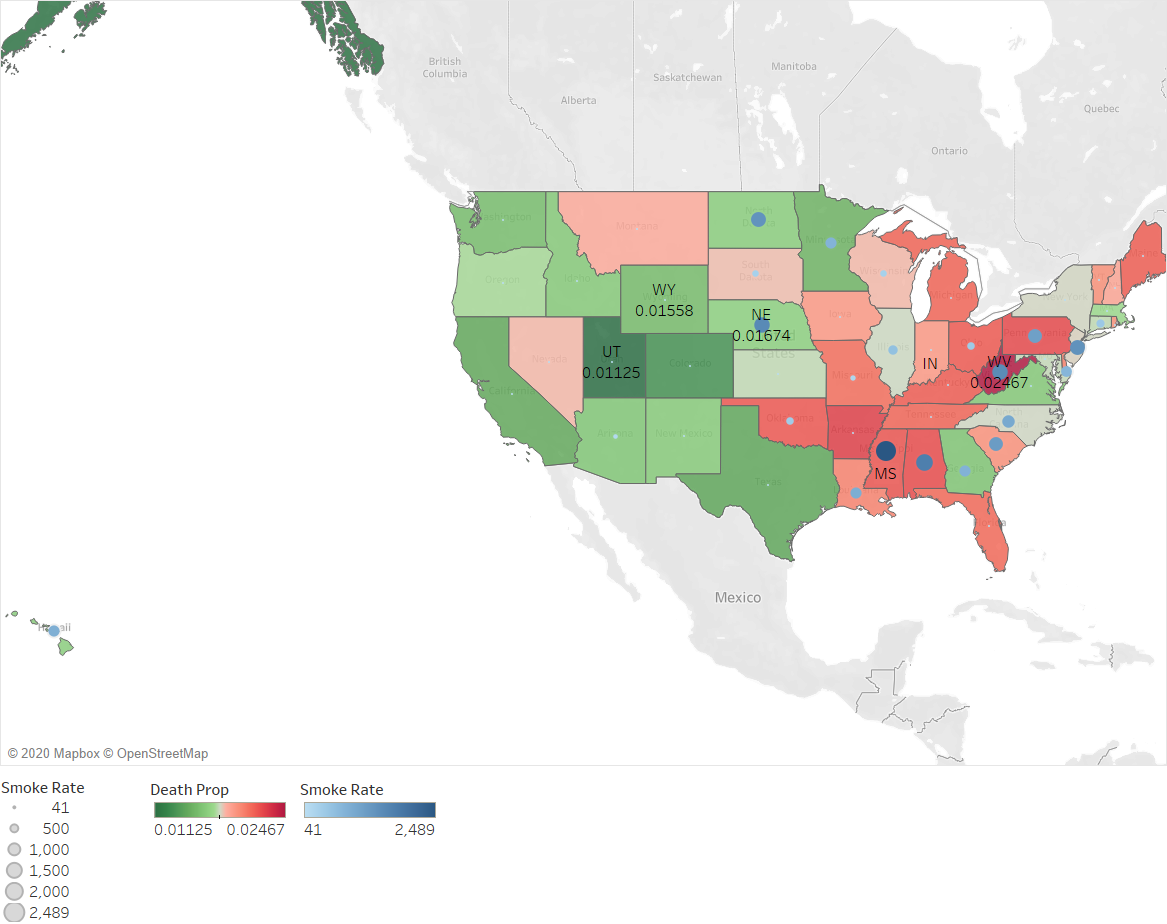
****

Figure18: Average death rate/smoke rate distribution heat map

A. This figure is a heat map showing the cumulative smoking rate and average rate of youth smoke-related death over the state's population from 2014 to 2017. The state which has the highest smoking death rate is West Virginia (0.616%), and lowest is Utah (0.281%).

B. And higher smoking rate (larger blue dots) is associated with higher death rate (more redness). Higher smoke rate is associated with higher youth smoke-related death rate.

C. The association has some outliers like Nebraska, might because of state-living or population structure difference: other reasons’ death rate rose a lot like obesity/Excessive drinking, decreased proportion of youth smoking-related death rate.

1. Scatterplot for the Smoke Rate and Mortality Rate in Recorded States During 2014-2017 - Qilin Liu

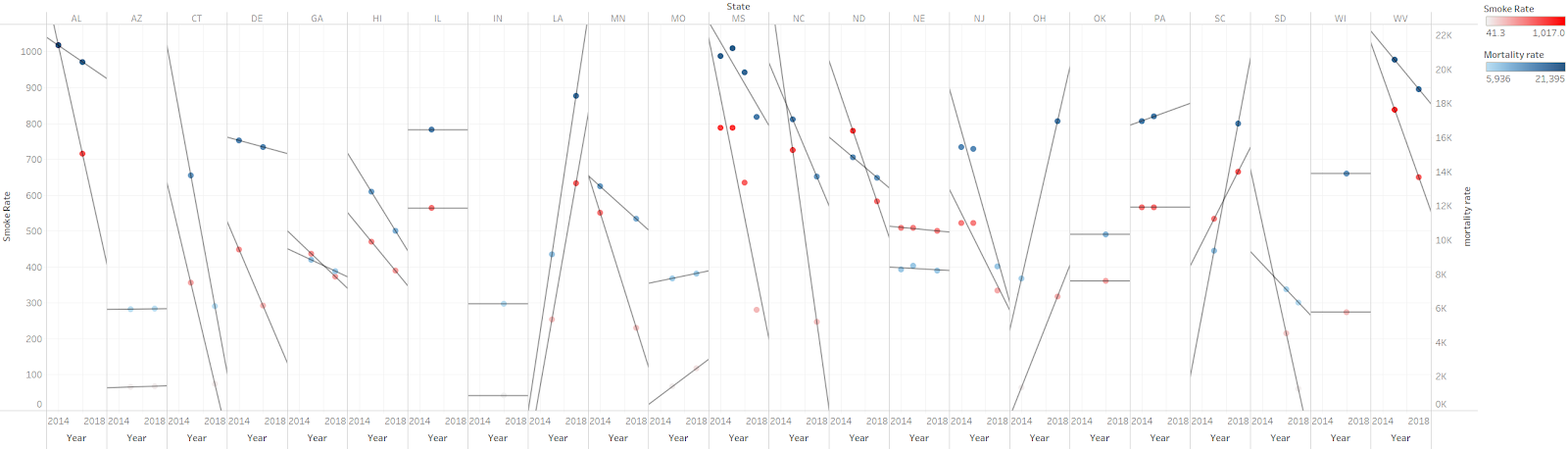


Figure19: scatterplot for the smoke rate and mortality rate in recorded states during 2014-2017

A. This figure is a scatterplot showing among all of the states, when the smoke rate rise/fall, the smoke-related mortality rate changes with the same trend.

B. It shows that the higher smoke rate is truly associated with higher smoke-related mortality rate.

C. The government and smoke-free organizations should propagate this fact to youth publics.

1. Horizontal Bar Chart for the Average of Smoke Rate Between Cessation-Willingness and Non-cessation Willingness Youth in Recorded States - Qilin Liu

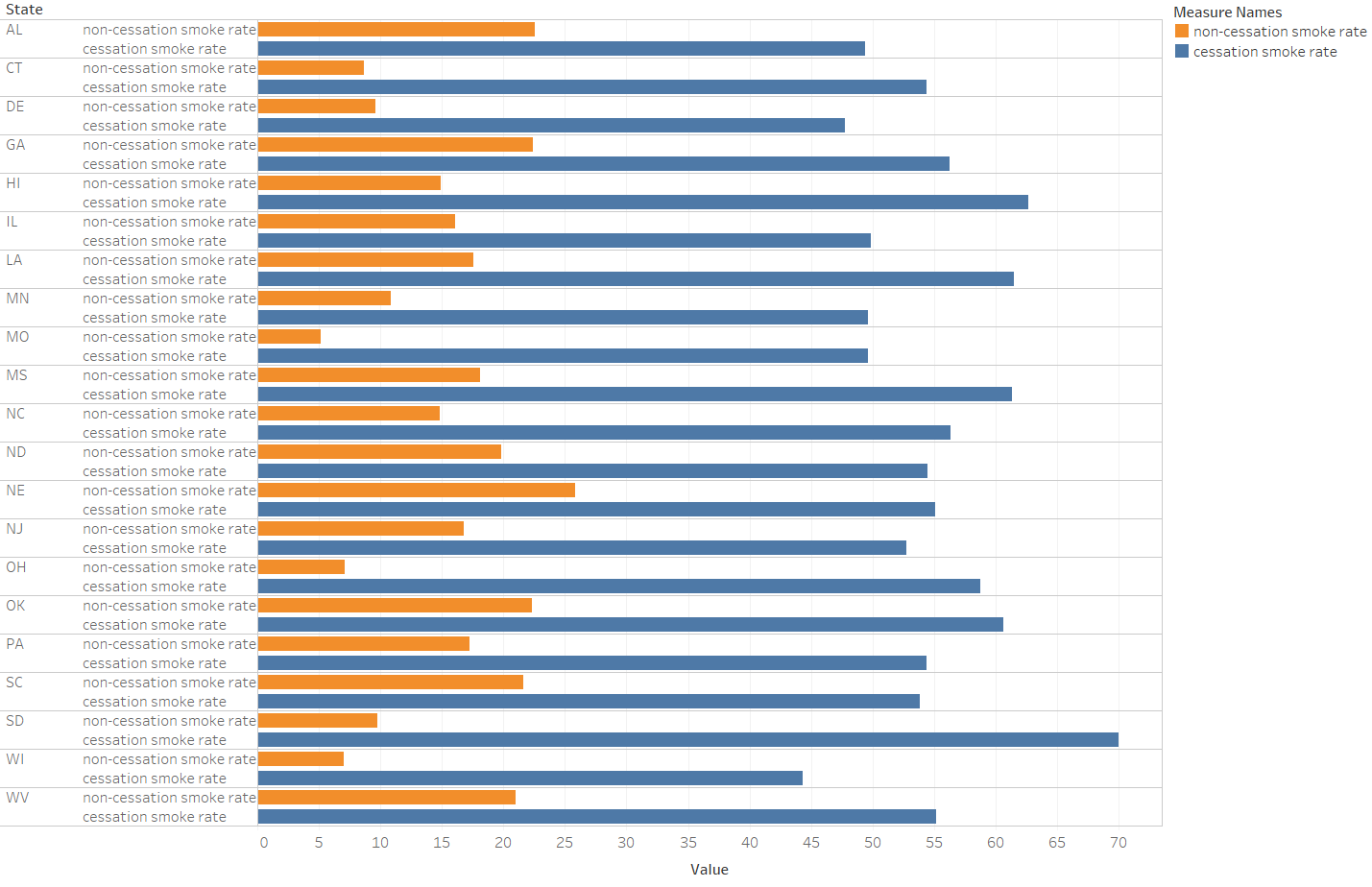


Figure20: horizontal bar for the average of smoke rate between cessation-willing and non-cessation willing youth in recorded states

A. This figure is a horizontal bar chart. It shows the smoke rate between cessation-willing and non-willing youth in recorded states.

B. All of the states show that the current youth smokers who want to quit have a higher smoke rate than the youth who are smoking while with no intention to cease smoking, whether cigarette or smokeless tobacco.

C. The government and smoke-free organizations should be alert that those youth who have no intention to quit, their smoke rate will rise higher until it is so serious that they have to cease it.

**C. Statistical Model: One Way ANOVA**

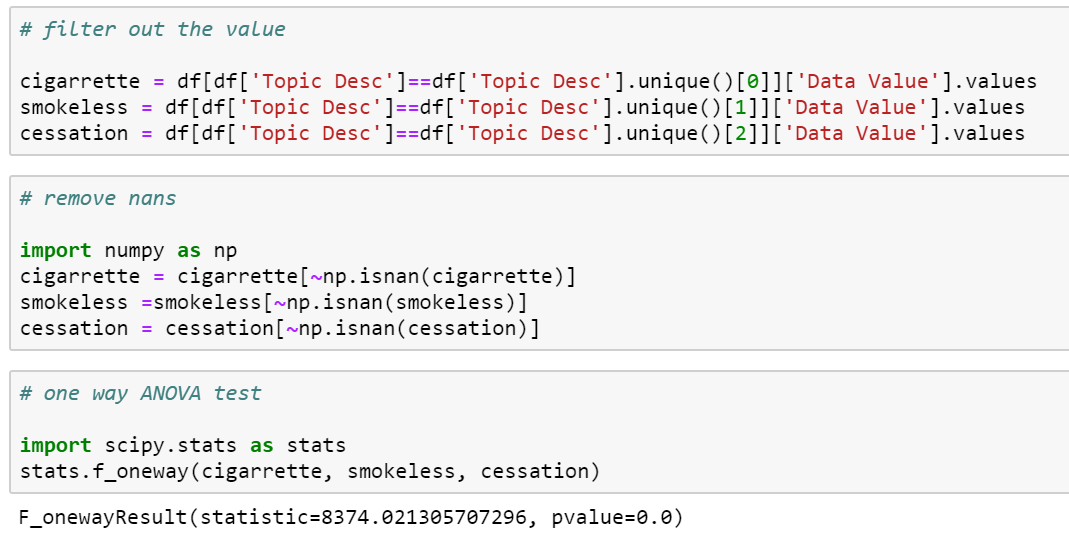
****

Figure21: Result of ANOVA for Various Smoking Types

**Hypothesis:** The smoking rates do not vary among different smoking types (cessation, cigarette, smokeless tobacco).

**A:** The figure indicates the result of the one way ANOVA test on if the smoking rates vary among different smoking types.

**B:** The stats is huge, and leads to a p-value close to 0, which gave us great confidence to reject the null hypothesis that there is no difference in smoking rate among smoking types.

**C:** This finding indicates that it would be beneficial to drill in different smoking types and make different customized policies to prevent smoking correspondingly.

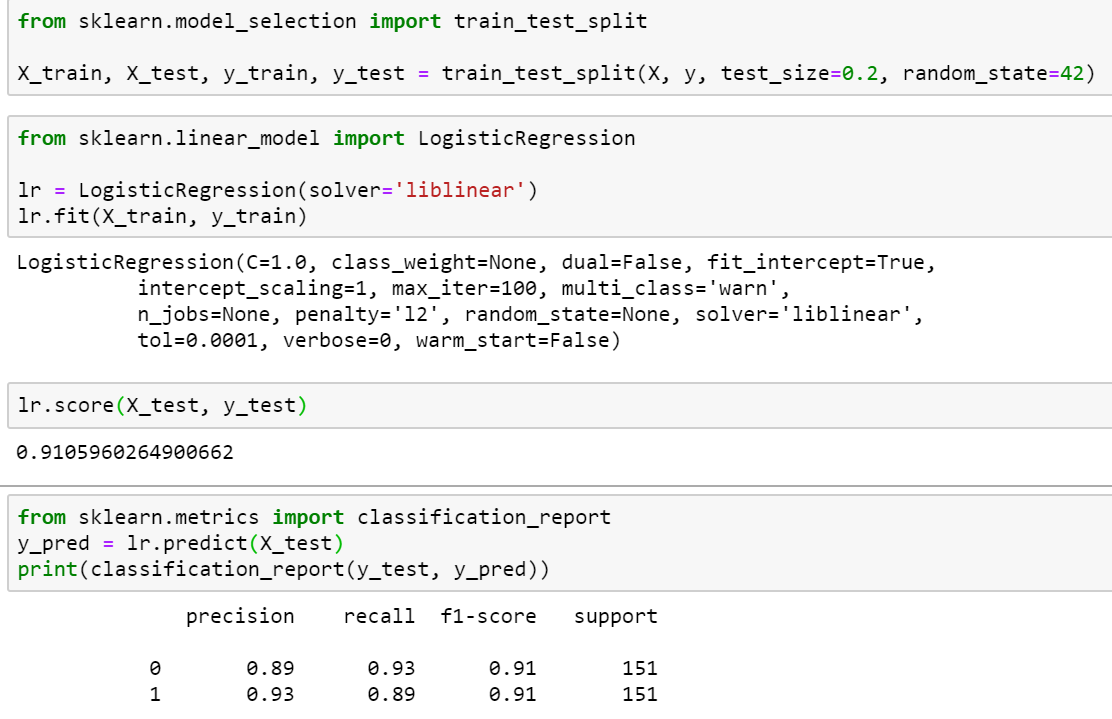
**D. Machine Learning: Logistic Regression**

Figure22: Logistic Regression for Smoking Rate Classification

**Hypothesis:** There exists a classification model that can accurately predict high or low cancer mortality rate with the youth smoking data.

**A:** Data is divided into train and test; where the dependent variable is cancer death rate. This figure showed the result of the classifier.

**B:** The performance indicated that the logistic regression could accurately predict whether a subgroup of the population would have a high or low cancer death rate.

**C:** The finding implied that knowing the youth smoking rate is helpful in predicting the cancer death rate.