Data Exploration

Changes in smoking and drinking rates in counties in each U.S. state and the number of related cancer cases from 2002 to 2012.

Student Name: PEIYU LIU
Student ID: 31153291
Tutor: Angel Das and Mohit Gupta
Lab: Tutorial 3, Thursdays at 4 pm6 pm

Table of content

1 Introduction Problem description, question, and motivation

Alcohol is becoming a popular primary drink when people do social or work activities. Alcohol abuse by people is a significant problem for personal health and public health. So I want to find out alcohol use increasing rate. I want to find out if the main groups of people who drink in the population are men, women, or both.

The poison in cigarette smoke weakens the body's immune system, making it harder to kill cancer cells. The poison in tobacco smoke damages or alters the DNA of the cells, which, when damaged, begins to grow out of control and form cancerous tumours (CDC). Alcohol is also a primary carcinogen, and all types of alcoholic beverages, including red and white wine, beer, cocktails and liquor, are associated with cancer. The more you drink, the higher your risk of cancer. So I also used cancers' case records to find out if drinking and smoking rates affected the number of cancer cases.

1. How did the alcohol consumption of the states in the United States change from 2002 to 2012? How do people who drink, drink heavy, and drink binge change? What are the trends in male and female drinking rates? Are men the main reason for drinking?

2. What is the trend in the percentage of smokers in the total population in each state in the United States from 2002 to 2012? What is the trend in the proportion of female smokers? Does the number of male and female smoking continue to rise? Which gender has the greater proportion of smoking increase/decrease?

3. How did the number of cancer-related cases in the United States change from 2002 to 2012? Are cancer cases on the rise? What was the usage rate of tobacco and alcohol in the same period? Is the hypothetical use rate of tobacco and alcohol-related to the number of instances?

2 Data Wrangling

R and Tableau are used in data wrangling and data checking. raw datasets contain four types of data: XML, XLSX, CSV and pdf. In detail, the "pdftools" package is used to do pdf data processing. "XML" package is used to do XML file processing. "readxl" package is used to do XLSX file processing.

```
# libraries
require("dplyr")
require("pdftools")
require("readxl")
require("XML")
require("methods")
require("tidyverse")
```

figure 1: libraries to import and process dataset

Data of "Heavy drinking and binge drinking rise sharply in US counties _ Institute for Health Metrics and Evaluation.pdf " is the alcohol drinking research summary in the USA, experts' conclusion can be used in support question 3.

Data of "Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, male 15+ years of age).xml " and "Total alcohol consumption per capita, male (liters of pure alcohol, projected estimates, male 15+ years of age).xml "can be used to analyse question 1. Data contains years distribution from 2000 to 2018, area distribution in the USA states, alcohol consumption for female and male populations older than 15 years.

Data of "IHME_USA_COUNTY_ALCOHOL_USE_PREVALENCE_2002_2012_NATIONAL_Y2015M04D23.X

LSX" contains year distribution from 2002 to 2012, gender distribution, alcohol drinking percentage changes from 2002 to 2012, country location distributions in the USA and all states distribution in the USA.
Data of "IHME_US_COUNTY_TOTAL_AND_DAILY_SMOKING_PREVALENCE_1996_2012.csv" contains genders, years from 1996 to 2012, the average number of cigarette usage rates in each state and county in the United States and rate boundaries.
Data of "https://gis.cdc.gov/Cancer/USCS/#/Trends/" contains all types of cancer population changes over time, the death count and cases' age in the states of USA. Only analyse cancers about Liver and Intrahepatic Bile Duct, Lung and Bronchus, Female Breast, Oral Cavity and Pharynx, Thyroid, which are most related to alcohol and cigarette.

```
require(XML)
require(methods)
#read female xml file
female_xml <- xmlParse('xml_alcohol/Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, male 15+ years of age).xml')
female_xml_df <- xmlToDataFrame(nodes = getNodeSet(female_xml,'//record'))
names(female_xml_df) <- c("USAcountry","description","year","consumption")
dim(female_xml_df)
# read male xml file
male_xml <- xmlParse('xml_alcohol/Total alcohol consumption per capita, male (liters of pure alcohol, projected estimates, male 15+ years of age).xml')
male_xml_df <- xmlToDataFrame(nodes = getNodeSet(male_xml,'//record'))
names(male_xml_df) <- c("USAcountry","description","year","consumption")
save(male_xml_df,file = 'male_xml_df.Rda')
save(female_xml_df,file = 'female_xml_df.Rda')

#read csv files: smoke csv file and csv cancer files
smoke_csv <-  read.csv('smoke/IHME_US_COUNTY_TOTAL_AND_DAILY_SMOKING_PREVALENCE_1996_2012.csv')
cancer_liver_bile_duct <- read.csv('cancer/ Liver and Intrahepatic Bile Duct.csv')
cancer_lung_bronchus <- read.csv('cancer/ Lung and Bronchus.csv')
cancer_female_breast <- read.csv('cancer/Female Breast.csv')
cancer_thyroid <- read.csv('cancer/Thyroid.csv')
cancer_oral_pharynx <- read.csv('cancer/Oral Cavity and Pharynx.csv')

# read xlsx file: alcohol xlsx files
# Reference: read xlsx with sheets:https://zhuanlan.zhihu.com/p/35608173
require(readxl)
alcohol_xlsx_any <- read_excel('alcohol/IHME_USA_COUNTY_ALCOHOL_USE_PREVALENCE_2002_2012_NATIONAL_Y2015M04D23.XLSX',
                           sheet = 2,range = NULL, col_names = TRUE,col_types = NULL, na = "", skip = 0, n_max = Inf)
alcohol_xlsx_heavy <- read_excel('alcohol/IHME_USA_COUNTY_ALCOHOL_USE_PREVALENCE_2002_2012_NATIONAL_Y2015M04D23.XLSX',
                             sheet = 3,range = NULL, col_names = TRUE,col_types = NULL, na = "", skip = 0, n_max = Inf)
alcohol_xlsx_binge <- read_excel('alcohol/IHME_USA_COUNTY_ALCOHOL_USE_PREVALENCE_2002_2012_NATIONAL_Y2015M04D23.XLSX',
                             sheet = 4,range = NULL, col_names = TRUE,col_types = NULL, na = "", skip = 0, n_max = Inf)
```

figure 2: Data wrangling and import

3 Data Checking

R and Tableau are used in data checking.

```
# clean data
female_xml_df_clean <- female_xml_df[which((female_xml_df$year > 2001)& (female_xml_df$year < 2013)),]
male_xml_df_clean <- male_xml_df[which((male_xml_df$year > 2001)& (male_xml_df$year < 2013)),]
cancer_lbd <- cancer_liver_bile_duct[which((cancer_liver_bile_duct$Year > 2001)& (cancer_liver_bile_duct$Year < 2013)),]
cancer_lb <- cancer_lung_bronchus[which((cancer_lung_bronchus$Year > 2001)& (cancer_lung_bronchus$Year < 2013)),]
cancer_fb <- cancer_female_breast[which((cancer_female_breast$Year > 2001)& (cancer_female_breast$Year < 2013)),]
cancer_td <- cancer_thyroid[which((cancer_thyroid$Year > 2001)& (cancer_thyroid$Year < 2013)),]
cancer_op <- cancer_oral_pharynx[which((cancer_oral_pharynx$Year > 2001)& (cancer_oral_pharynx$Year < 2013)),]
smoke_clean <- smoke_csv[which((smoke_csv$year > 2001)& (smoke_csv$year < 2013)),]
alcohol_any_clean <- alcohol_xlsx_any
alcohol_heavy_clean <- alcohol_xlsx_heavy
alcohol_binge_clean <-alcohol_xlsx_binge

# save cleaned data
write.csv(female_xml_df_clean,file = 'female_xml_df_clean.csv')
write.csv(male_xml_df_clean,file ='male_xml_df_clean.csv')
write.csv(cancer_lbd,file ='cancer_liver_bile_duct.csv')
write.csv(cancer_lb,file ='cancer_lung_bronchus.csv')
write.csv(cancer_fb,file ='cancer_female_breast.csv')
write.csv(cancer_td,file ='cancer_thyroid.csv')
write.csv(cancer_op,file ='cancer_oral_pharynx.csv')
write.csv(smoke_clean,file ='smoke_csv.csv')
write.csv(alcohol_any_clean,file ='alcohol_any_clean.csv')
write.csv(alcohol_heavy_clean,file ='alcohol_heavy_clean.csv')
write.csv(alcohol_binge_clean,file ='alcohol_binge_clean.csv')
```

figure 3: filter year period

Read the data into r studio, then view the data content through the data cache in the environment, use dim() to view the specific record number and structure of the data, and use names() to view the variable name of the data. Import the data into Tableau to view the data structure and data null values. It is determined that null values exist and cannot be deleted. The reason is that data statistics are not performed in some areas of the United States. Data exploration aims to analyse data in the 2002 to 2012 period. So that data filtering will be done before data visualisation.

4 Data Exploration

Use Tableau to do initial data exploration (D3 and R will be used in final data visulisation).

4.1 Data exploration for question 1(mentioned in 1 part)

There are three levels of alcoholism: "Any", "Heavy", and "Binge". "Any" drinking is defined as at least one drink of any alcoholic beverage in the past 30 days. "Heavy" drinking is defined as the consumption, on average, of more than one drink per day for women or two drinks per day for men in the past 30 days. "Binge" drinking is defined as the consumption of more than four drinks for women or five drinks for men on a single occasion at least once in the past 30 days.

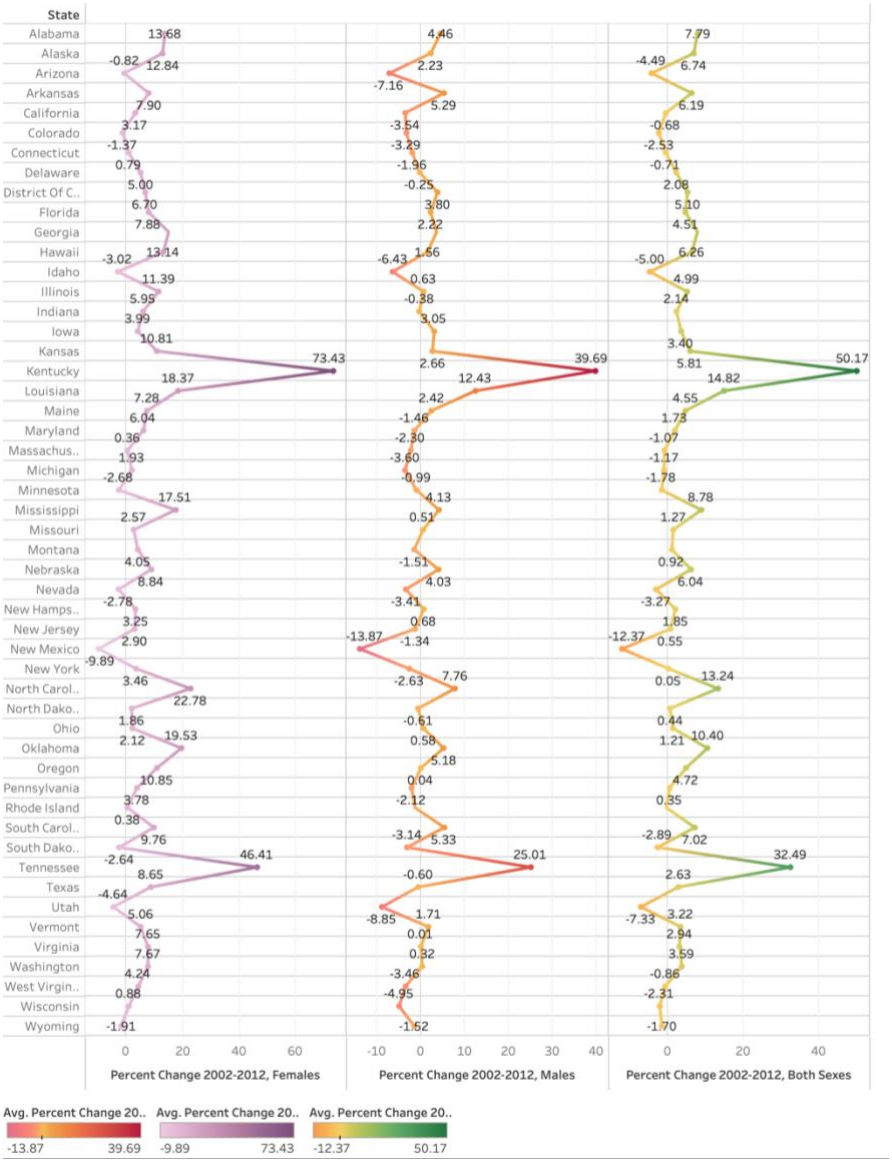

figure 4: Drink rate trend(Any level)(link to the original figure 4 with Monash account))

For question 1. The overall drinking trend of all states in the United States from 2002 to 2012 (show in figure 4). Import the processed drinking statistics of the states in the United States into Tableau, classify the groups by region using states as the parameter, and use gender to generate gender classification for the groups. Drag the average drinking rate data into it to visualize the data. It can be seen that from 2002 to 2012, the drinking rate of all states in the United States remained flat with a slight increase. In some areas, such as Kentucky and Tennessee, the drinking rate has increased significantly, and the drinking rate of women in Kentucky has risen sharply or even by 73. It can be inferred that the rising trend of female drinking rates has led to an increase in the overall drinking rate of the region. The 10-year changes in women's drinking rates in various states (show in figure 5) and the changes in women's drinking rates in the past ten years are drawn using the state as a classification standard, and they are generally on the rise.

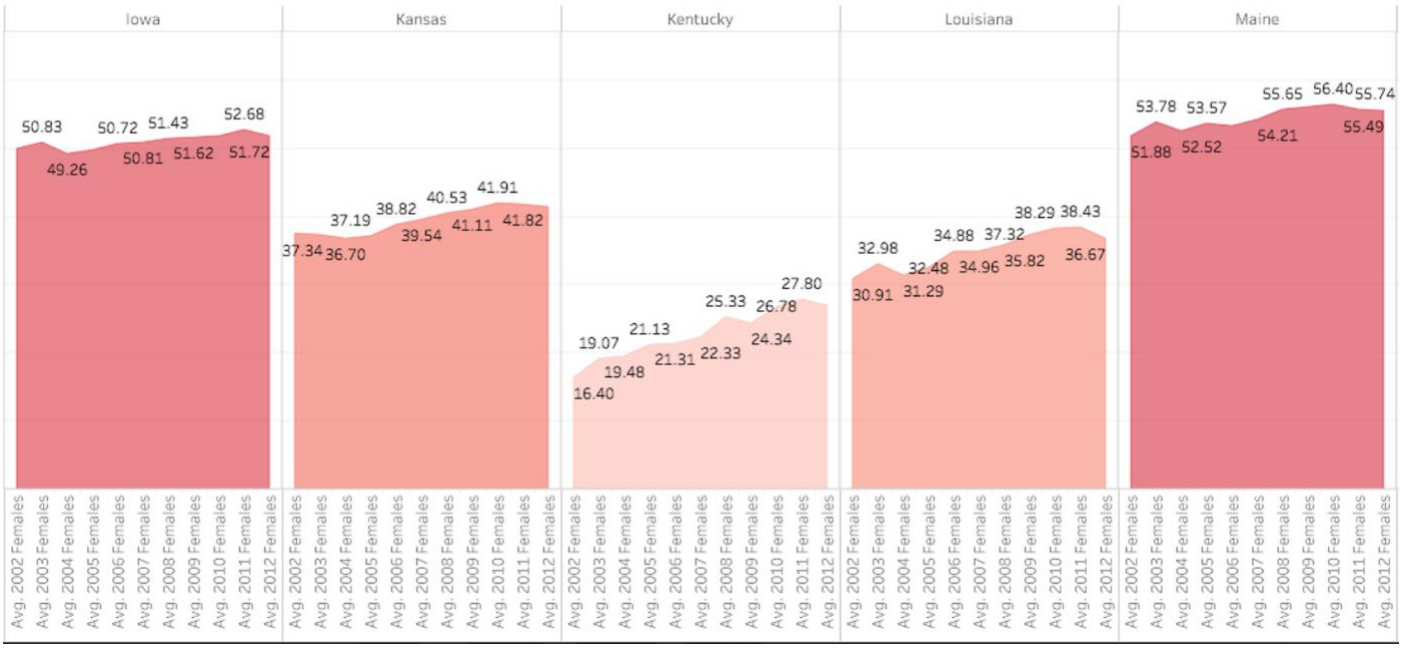

figure 5：Part of female drinking rate(link to the original figure 5 with Monash account)

The same analysis method also applies to Heavy level(show in figure 6, link to original figure 6 with Monash account) and Binge level(show in figure 7, link to original figure 7 with Monash account) of alcohol consumption. The same analysis method is also applicable to the Heavy level and Binge level of alcohol consumption. According to the group by the states, the visualization shows that although the people who drink Heavy and Binge levels are less than the Any level, the general trend is also on the rise, with an increase in the rate of alcohol abuse. But after 2011, the rate of heavy drinkers began to decline, with a marked downward trend. Binge level drinkers are still on the rise in the most regional states.

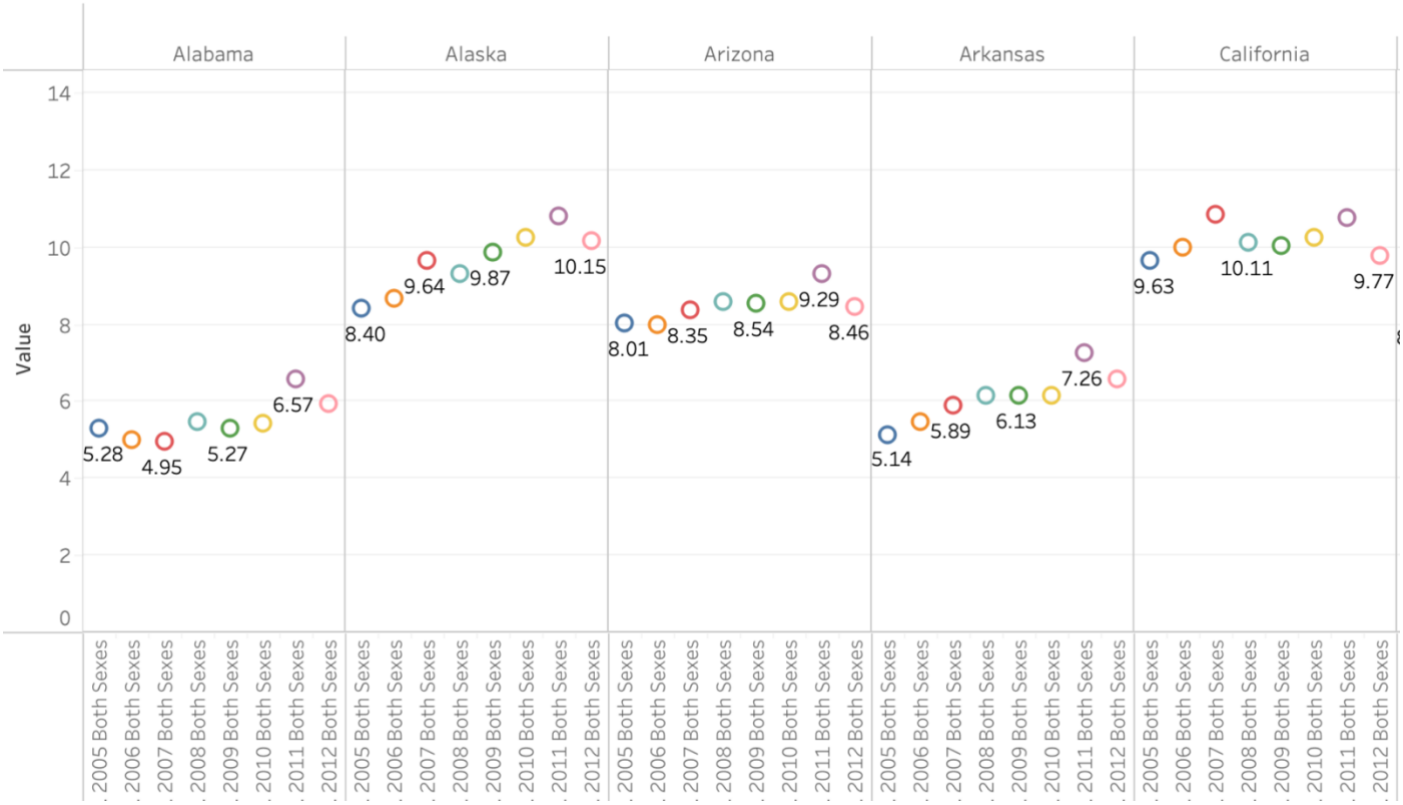

figure 6：Part of drink heavy trend (link to original figure 6 with Monash account)

figure 7：Part of drink Binge trend (link to original figure 7 with Monash account)

4.2 Data exploration for question 2(mentioned in 1 part)

For question 2, use checked smoke statistics data to explore data after filtering and processing data in R studio. Use cleaned new smoke data into Tableau to do visualization. Question 2 is aimed to figure out smoking rate changes from 2002 to 2012 in the USA. Data is the group by males and females, and years are as time distribution, states in the USA as location distribution. Use these conditions to get the figure (shown in figure 8) to analyze question 2. the analysis result is for most even each state in the USA, the male is the main population who is smoking. The male smoking rate is always high than females. The smoking rate changes

between 10 years, although the rise and fall trend is existing, the whole trend is that the smoking rate is decreasing in most states (show in figure 9).

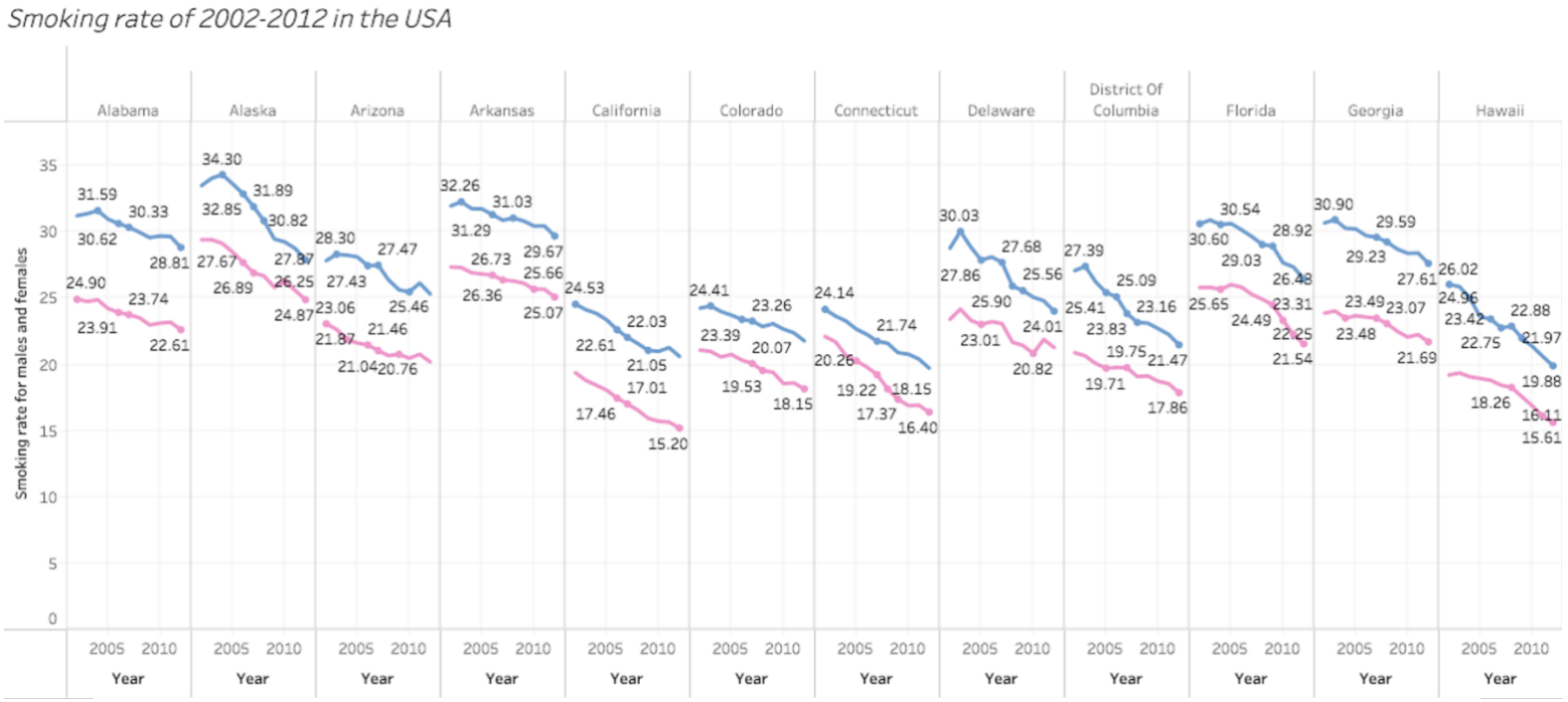

figure 8:Part of Smoking rate changes (link to original figure 8 with Monash account)

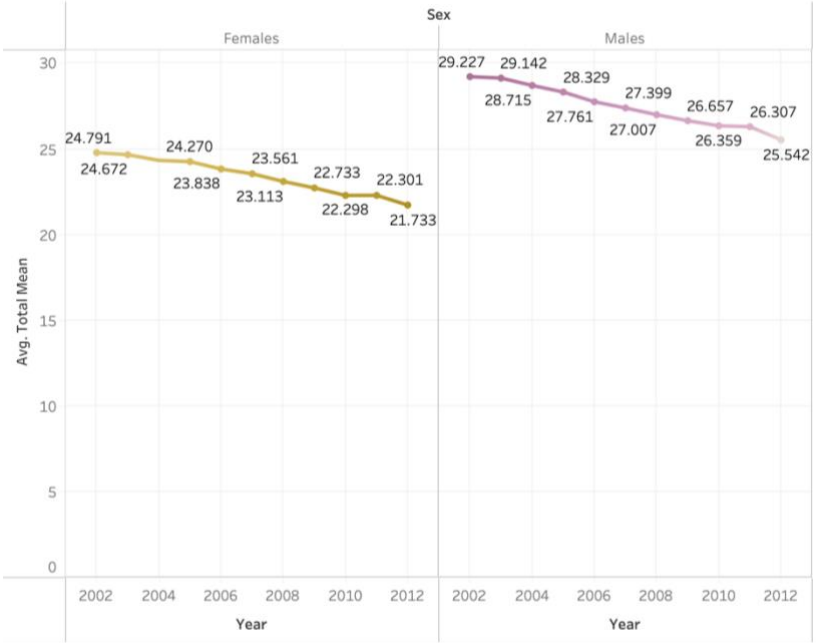

figure 9: Smoking rate trend with genders (link to original figure 9 with Monash account)

4.3 Data exploration for question 3(mentioned in 1 part)

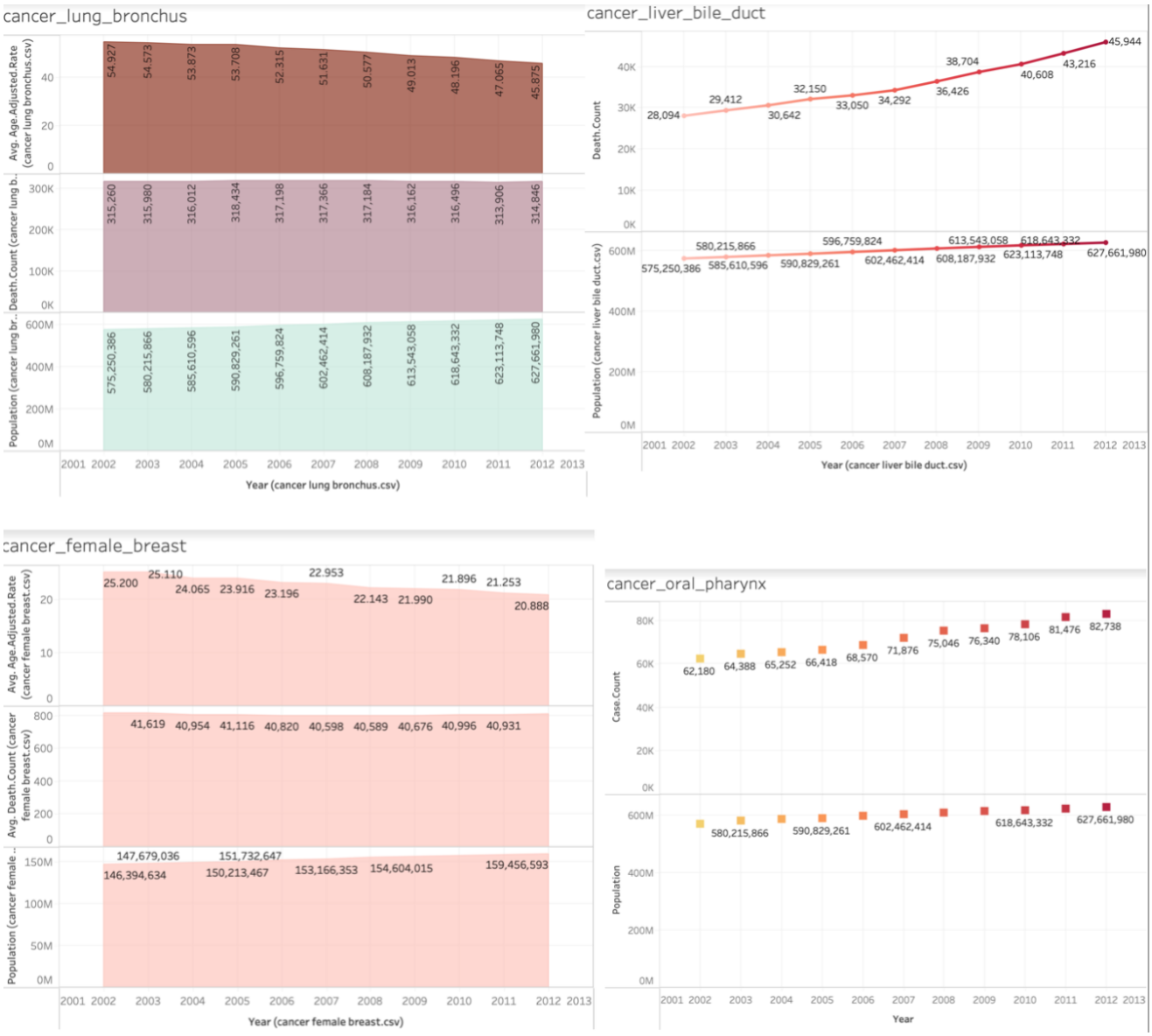

figure 10: Breast, liver, lung, oral cancers (link to original figures with Monash account)

Breast, liver, lung, oral, and thyroid cancers are closely related to alcohol and tobacco, and cancer is positively correlated with carcinogens. The more carcinogens people consume, the greater your chances of getting cancer. In areas with high tobacco and alcohol use, cancer incidence is higher, and the population value is higher. By importing various types of cancer data, the group is grouped with the time distribution of yes, the cancer case, the cancer population, and the cancer age. Breast, liver, lung, oral cancers' cases are increasing with the years significantly (show in figure 10). Especially the death cases of liver cancer are growing dramatically, and the one reason might be the increasing alcohol drinking rate of males. The average age of people with lung cancer is becoming lower, the one reason might be smoking becoming popular, and in some states, the smoking rate increased, so that more young people try to smoke. Meanwhile, oral cancer cases are increasing fast, and death cases also growing. It might be related to the increasing trend of alcohol use in states.

Alcohol and cigarette will cause thyroid cancer, and this is the most common cancer in nowadays world. The main reason is because of people's diet. The whole trend in stats in the USA (show in figure 11). the population of people with thyroid is increasing, although not significantly, but the number also has nearly 2,500,000 increasing. The death cases due to thyroid cancer in the USA is increasing evidently, from 24058 to 47518. the reason might be that thyroid cancer is prevalent cancer. In the initial stage, people will not feel much uncomfortable, so people ignore periodic inspection. In daily life, they still use alcohol and tobacco.

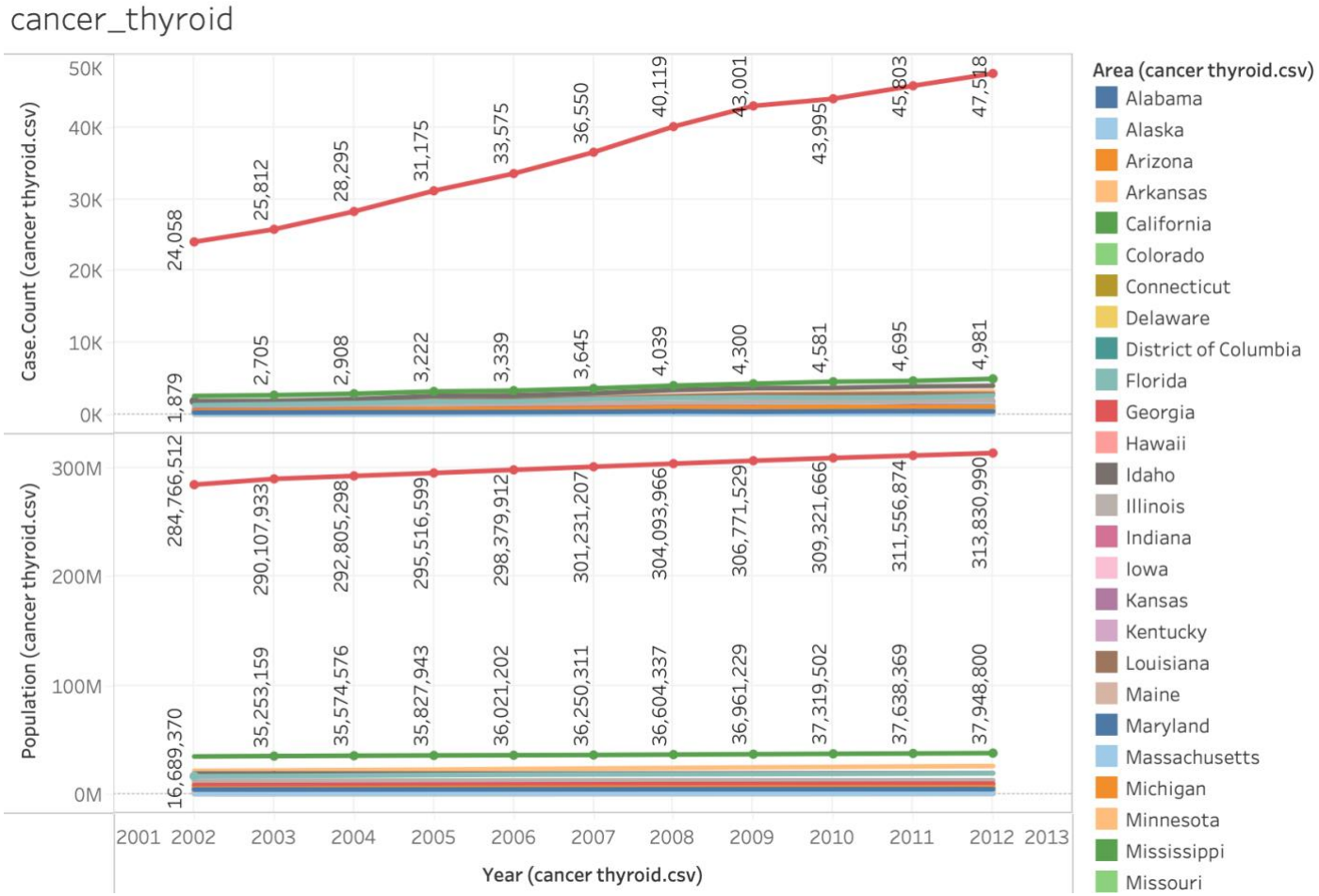

figure 11：Thyroid cancer trend [(link to original figure 11 with Monash account)]

5 Conclusion

The data covers all states and counties in the United States. From the visualization, it can be seen that the drinking rate of each state is in a slight upward trend, and the drinking rate of individual states has increased significantly. For example, Kentucky has risen by more than 50%. The overall trend of smoking rates in various states from 2002 to 2012 was a slight decline, but during the period 2002 to 2006, the smoking rate of women increased, but the smoking rate of men was always higher than that of women and men were the leading smoking group. Between 2002 and 2012, the number of women with breast cancer showed a clear upward trend, which may be related to the increase in the number of women who smoke and drink. The number of patients with liver cancer and oral cancer continues to rise, and the number of deaths from liver cancer has almost doubled in 10 years. This is closely related to the amount of alcohol consumed by men and women. The continuous increase in the number of people who have lung cancer and the decrease in the average age of the disease is related to the decrease in the average age of the smoking population. The number of people with thyroid cancer in each state has shown a slight upward trend. The overall number of people with thyroid cancer

in the United States has increased, and the number of deaths has risen sharply. Alcohol is also the leading cause of thyroid cancer.

6 Reflection

This task is mainly to use tools like Tableau and R for data analysis. The data format and content are relatively standardized. It is not difficult to clean and organize. It Only needs to filter out the unnecessary data and save the CSV format file for reading. It tries to analyze the problem from multiple angles based on the available data set. Epidemic data involves all states and counties in the United States, so it is very troublesome to classify through more than 3,000 counties, and the visualization is not intuitive enough. As for drawing graphics, too many data blocks are classified by state, and the pictures in Tableau cannot be well inserted into the report. There will be counties with the same name in each state in the United States, which leads to duplication when Tableau automatically generates latitude and longitude. Before making visualization, the name of each state should be replaced with latitude and longitude in the data collation process. Using the map drawing package to map the locations and data for each county is better; it is more intuitive and beautiful, and I will do it in the final visualization.

Bibliography

Heavy drinking and binge drinking rise sharply in US counties: http://www.healthdata.org/news-release/heavy-drinking-and-binge-drinking-rise-sharply-us-counties

Alcohol and Cancer Risk: https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet

Bagnardi, V., Rota, M., Botteri, E., Tramacere, I., Islami, F., Fedirko, V., Scotti, L., Jenab, M., Turati, F., Pasquali, E., Pelucchi, C., Galeone, C., Bellocco, R., Negri, E., Corrao, G., Boffetta, P., & La Vecchia, C. (2015). Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta-analysis. *British journal of cancer*, *112*(3), 580–593. https://doi.org/10.1038/bjc.2014.579

Inoue-Choi, M., Liao, L. M., Reyes-Guzman, C., Hartge, P., Caporaso, N., & Freedman, N. D. (2017). Association of Long-term, Low-Intensity Smoking With All-Cause and Cause-Specific Mortality in the National Institutes of Health-AARP Diet and Health Study. *JAMA internal medicine*, *177*(1), 87–95. https://doi.org/10.1001/jamainternmed.2016.7511

Nelson, D. E., Jarman, D. W., Rehm, J., Greenfield, T. K., Rey, G., Kerr, W. C., Miller, P., Shield, K. D., Ye, Y., & Naimi, T. S. (2013). Alcohol-attributable cancer deaths and years of potential life lost in the United States. *American journal of public health*, *103*(4), 641–648. https://doi.org/10.2105/AJPH.2012.301199

Chen, W. Y., Rosner, B., Hankinson, S. E., Colditz, G. A., & Willett, W. C. (2011). Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk. *JAMA*, *306*(17), 1884–1890. https://doi.org/10.1001/jama.2011.1590

Appendix

Original data exploration figures: https://drive.google.com/drive/folders/1Y1HGtOK_2pjjCkc5TN-kLebatY5tty_7?usp=sharing

Data wrangling and checking code: https://drive.google.com/file/d/1qZsAgMgbyrWq0cAV6_ksj7t7XfibFgKo/view?usp=sharing

Raw dataset: https://drive.google.com/file/d/16psO_3dhY26n5rej4iEYJrh0_q2y1VK4/view?usp=sharing