# Dataset Exploration Report

**Shengjue Yuan/ Tony Wang**

Warm up

- -Dataset reintroduction
- -Dataset exploration

Data preprocessing

- -Dataset cleaning

Dataset visualization

- -visualization of each variable in each question
- -visualization of each question with ggmap

# WARM UP

Dataset reintroduction

Dataset exploration

# Dataset reintroduction

Data: behavioral risk factor data of tobacco use

Observation: We have 38051 observations and each obs represents a percentage of a group of people that are of same age, same education level, gender, race with the same answer of a question asked in a survey

Observation example:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | YEAR | LocationA | LocationD | TopicType | TopicDesc | MeasureD | DataSourc | Response | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Low_Conf | High_Conf | Sample_Si | Gender | Race | Age | Education |
| 2 | 2010 | AL | Alabama | Tobacco U | Cessation | Quit Atten | BRFSS | | % | Percentag | 53.3 | | | 2.6 | 48.2 | 58.4 | 659 | Female | All Races | All Ages | All Grades |
| 3 | 2010 | AL | Alabama | Tobacco U | Cigarette U | Current Sn | BRFSS | | % | Percentag | 18.7 | | | 0.8 | 17.2 | 20.2 | 5234 | Female | All Races | All Ages | All Grades |
| 4 | 2010 | AL | Alabama | Tobacco U | Cigarette U | Current Sn | BRFSS | | % | Percentag | 18.6 | | | 1.4 | 15.9 | 21.3 | 1197 | Female | All Races | 18 to 44 Y | All Grades |

# DATASET EXPLORATION

```
levels(Tobacco$TopicDesc)
#[1] "Cessation (Adults)"            "Cigarette Consumption (Adults)"
#[3] "Cigarette Use (Adults)"

#1.get rid of Cessation (Adults) and Cigarette Consumption (Adults)
Tobacco.CU<-Tobacco[Tobacco[,5]!="Cessation (Adults)" &  Tobacco[,5]!="Cigarette Consumption (Adults)",]

#2.get rid of useless columns
Tobacco.CU1<-Tobacco.CU[,-c(3,4,5,6,7,9,10,12,13,14,15,16,18,19,20,21,23,24)]

#3.Explore the data set, we find out each state in each year need to answer 3 questions and
# the sum of the percentage among the responses for each question would be the 100 percent
Tobacco.CU1[Tobacco.CU1[,1]==2000 & Tobacco.CU1[,9]=="8AGE" & Tobacco.CU1[,8]=="1GEN" & Tobacco.CU1[,10]=="6RAC" & Tobacco.CU1[,11]=="6EDU" &  Tobacco.CU1[,2]=="AL"
# YEAR LocationAbbr  Response Data_Value Sample_Size                          GeoLocation MeasureId StratificationID1 StratificationID2 StratificationID3
# 501 2000           AL                25.2        2234 (32.84057112200048, -86.63186076199969)    110CSA             1GEN             8AGE             6RAC
# 515 2000           AL Every Day      76.4         549 (32.84057112200048, -86.63186076199969)    166SSP             1GEN             8AGE             6RAC
# 519 2000           AL Some Days      23.6         549 (32.84057112200048, -86.63186076199969)    166SSP             1GEN             8AGE             6RAC
# 521 2000           AL   Current      25.2        2234 (32.84057112200048, -86.63186076199969)    165SSA             1GEN             8AGE             6RAC
# 524 2000           AL    Former      24.0        2234 (32.84057112200048, -86.63186076199969)    165SSA             1GEN             8AGE             6RAC
# 527 2000           AL     Never      50.8        2234 (32.84057112200048, -86.63186076199969)    165SSA             1GEN             8AGE             6RAC
# StratificationID4
# 501              6EDU
# 515              6EDU
# 519              6EDU
# 521              6EDU
# 524              6EDU
# 527              6EDU
```

3 topics
3 questions in Cigarette Consumption

# Data preprocessing

Data cleaning

Creating new dataset

# Data Cleaning/Processing

Step 1. Get rid of NAs.

Step 2. Discard topics that we are not exploring and columns that are useless.

Step 3. Separate the data into 3 data frames according to the questions they are asked.

Step 4. Delete gender, age and education for frequency and status dataset because they are constant.

# Topics to be analyzed

As we mentioned in our snap talk, our dataset includes following topics:

```
> levels(data$MeasureDesc)
[1] "Current Smoking"
[2] "Current Smoking  (2 yrs Race/Ethnicity)”
[3] "Daily Cigarette Consumption Among Everyday Smokers - Average"
[4] "Daily Cigarette Consumption Among Everyday Smokers - Frequency Categories"
[5] "Percent of Former Smokers Among Ever Smokers"
[6] "Quit Attempt in Past Year Among Everyday Cigarette Smokers"
[7] "Smoking Frequency"
[8] "Smoking Status"
```

# Create new dataset

step 1:
three new dataset that each dataset represent all the answers for each questions

step 2:
a new dataset that each row represent one state in one year containing all the responses for all three questions

| Year | new.place | cu | FEM | FEF | FSM | FSF | SCM | SCF | SFM | SFF | SNM | SNF | SCA | SFA | SNA | FEA | FSA | long | lat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1996 | Alabama | 22.4 | 74.8 | 82.2 | 25.2 | 17.8 | 24.3 | 20.7 | 29.3 | 14.6 | 46.4 | 64.7 | 22.4 | 21.5 | 56.1 | 78.5 | 21.5 | -86.90230 | 32.31823 |
| 1997 | Alabama | 24.6 | 86.0 | 82.1 | 14.0 | 17.9 | 28.6 | 21.2 | 29.6 | 15.3 | 41.8 | 63.5 | 24.6 | 22.0 | 53.3 | 84.2 | 15.8 | -86.90230 | 32.31823 |
| 1998 | Alabama | 24.6 | 78.7 | 78.8 | 21.3 | 21.2 | 27.2 | 22.3 | 28.9 | 14.9 | 43.8 | 62.8 | 24.6 | 21.5 | 53.9 | 78.8 | 21.2 | -86.90230 | 32.31823 |

# Data Visualization

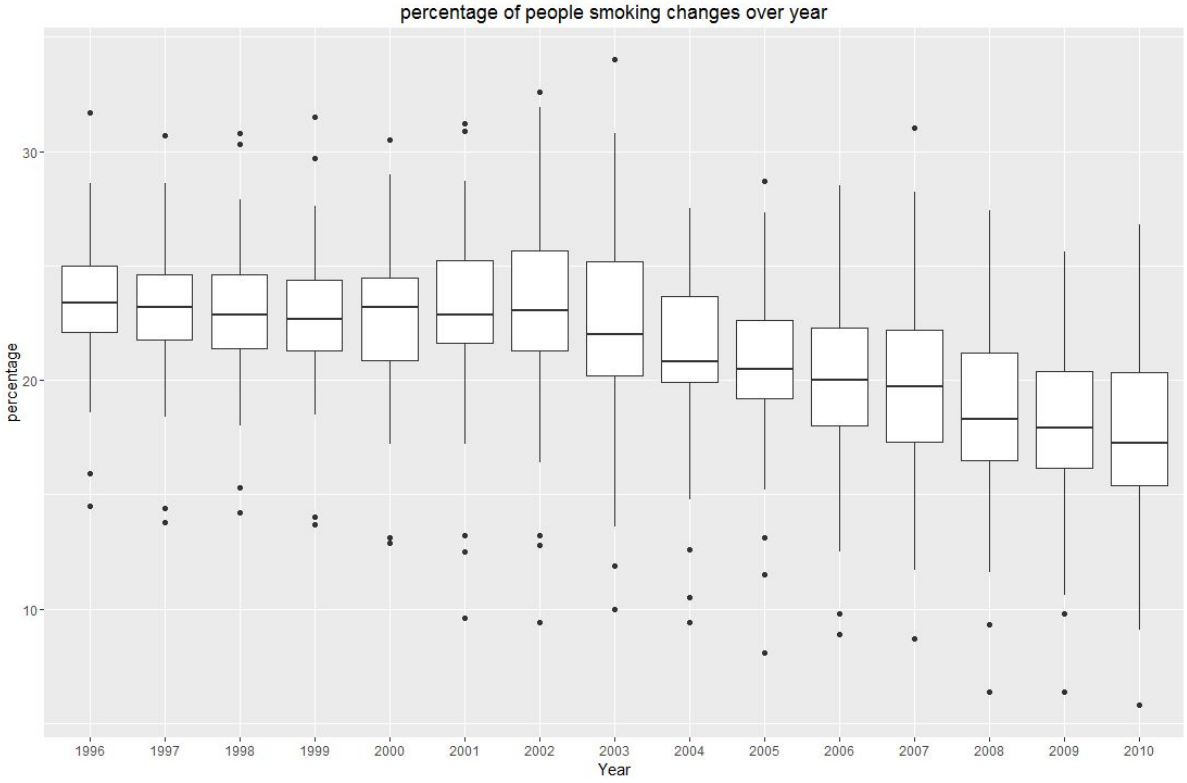
First we are going to analyze current smoking dataset.

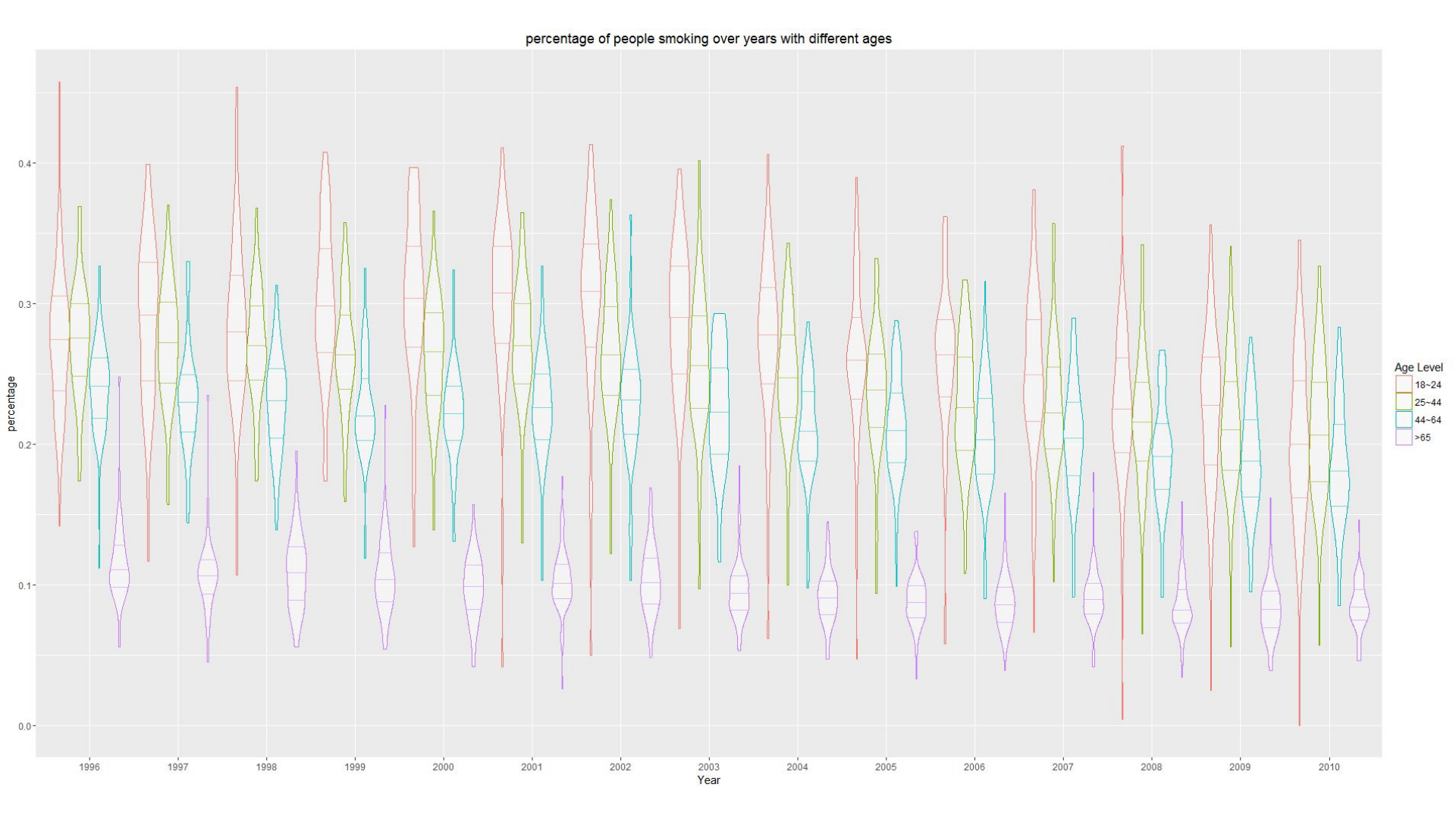
This dataset measures the percentage of current smokers among all people investigated.

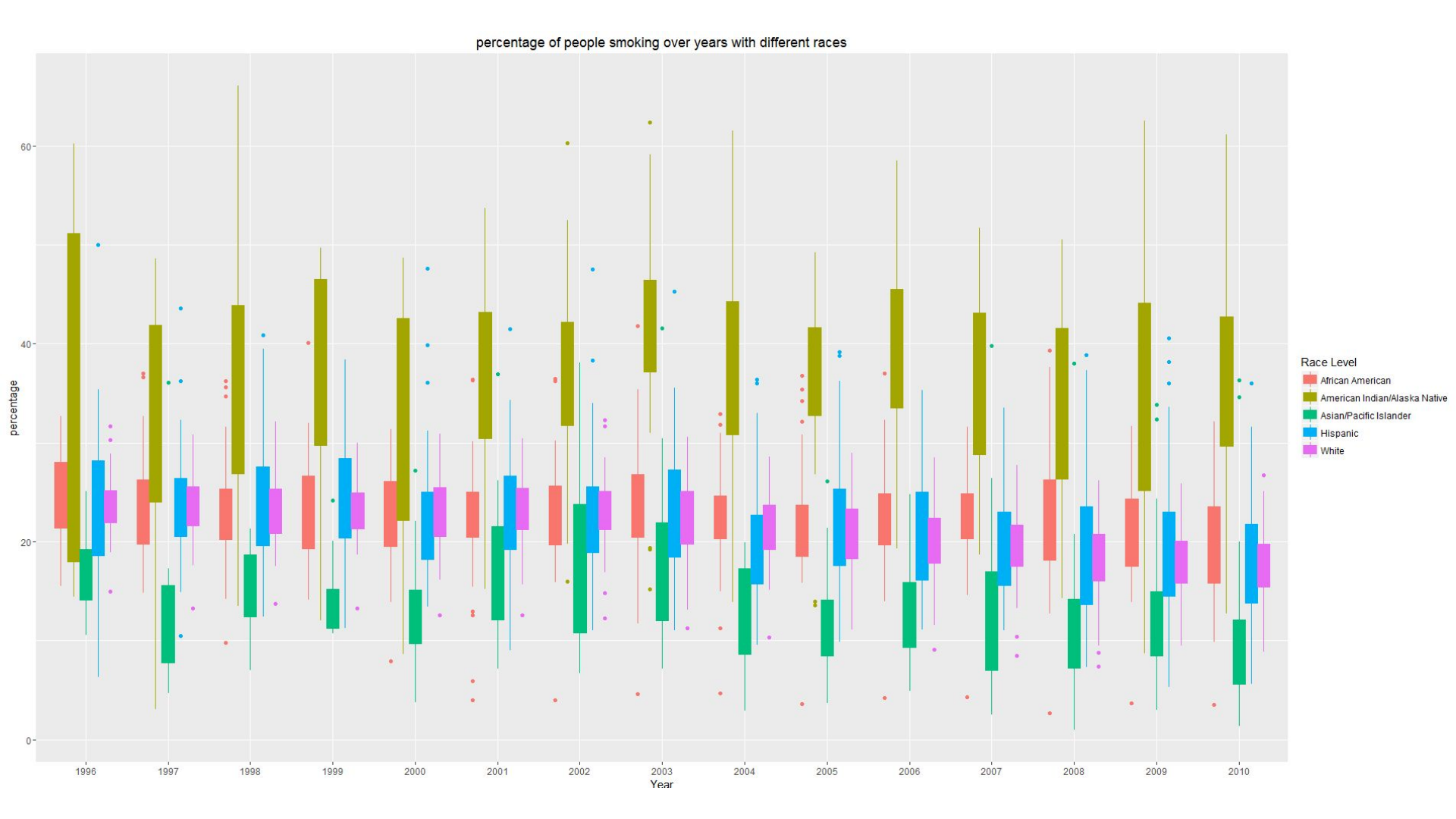
The data value is dependent by year, location, gender, age, race and education. So we made a plot for each variable.

# Variable: Year

The dataset includes information from 1996 to 2010, we represent the relationship by boxplot:




percentage of people smoking over years with different ages
percentage
0.0
0.1
0.2
0.3
0.4
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year
Age Level
18~24
25~44
44~64
>65

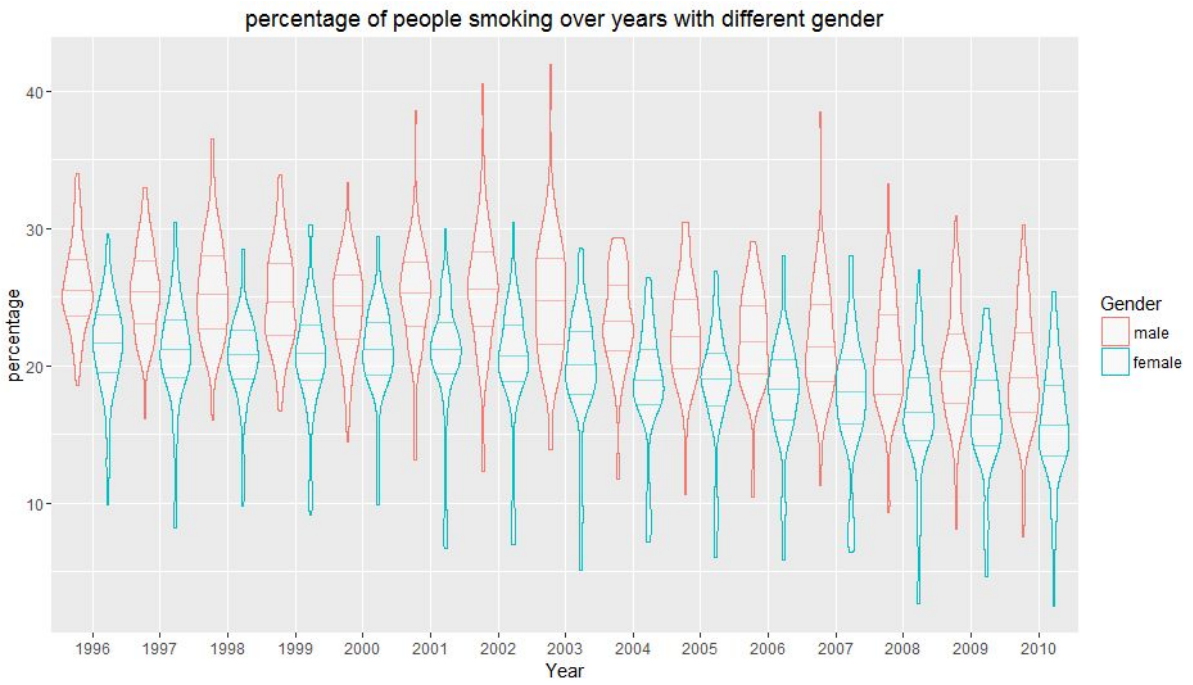

percentage of people smoking over years with different races
percentage
60
40
20
0
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year
Race Level
African American
American Indian/Alaska Native
Asian/Pacific Islander
Hispanic
White

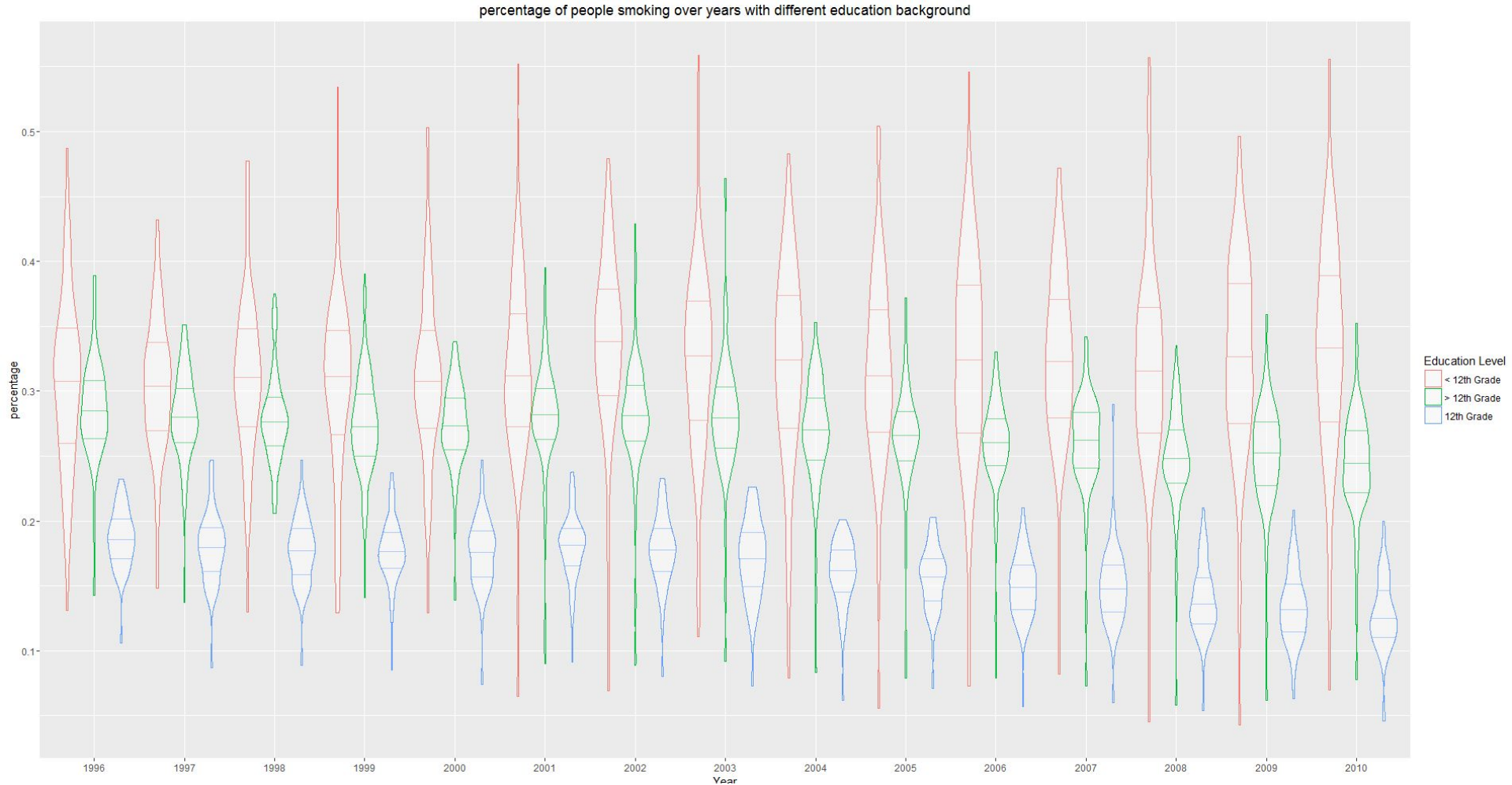
# Variable: Gender

Genders include:

```
levels(Tobacco.CU[,"Gender"])
# [1] "Female"  "Male"    "Overall"
```

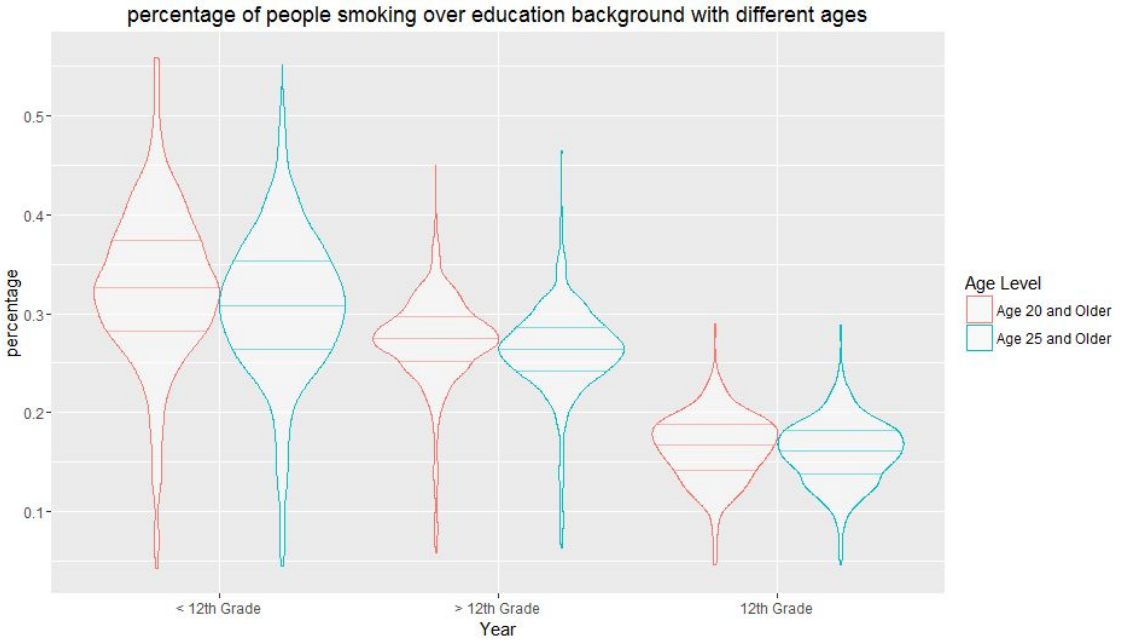
Simply deleting “Overall” observation will provide us with a graph showing difference of smoking status between male and female.




percentage of people smoking over years with different education background
percentage
0.5
0.4
0.3
0.2
0.1
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year
Education Level
< 12th Grade
> 12th Grade
12th Grade

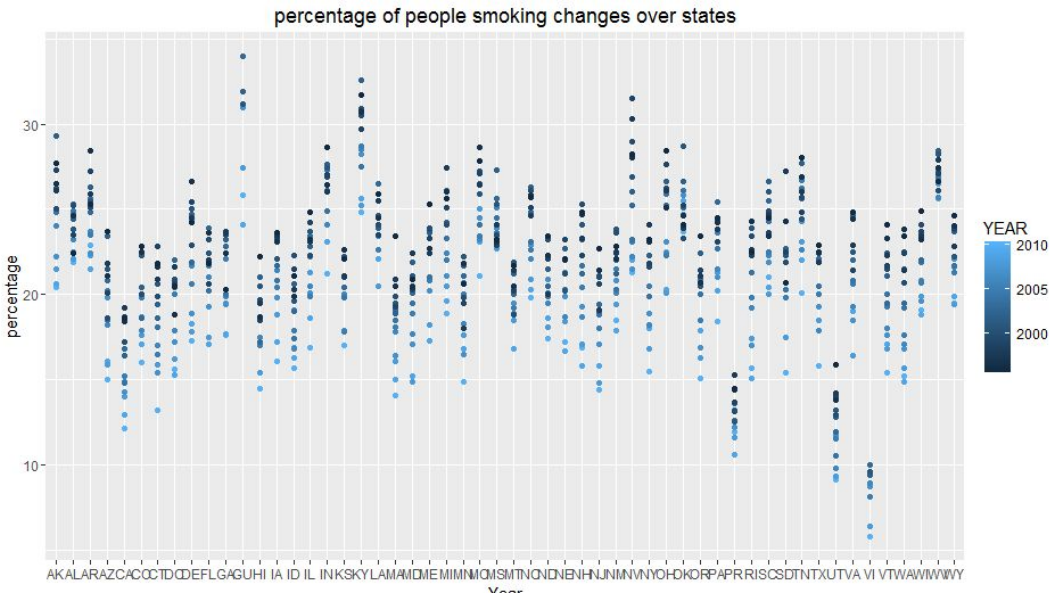
# Variable: Education

Before, we mentioned that “Age 20 and older” and “Age 25 and older” will only appear when education is not “All Grades”. Then how do these two variables related? We create a violin plot about this.



# Variable: Location(state)

There are 55 states in U. S. and creating a plot with a categorical variable with 55 levels will be very difficult. We consider location as a quantitative variable and draw a scatterplot to show the distribution.



# Smoking frequency

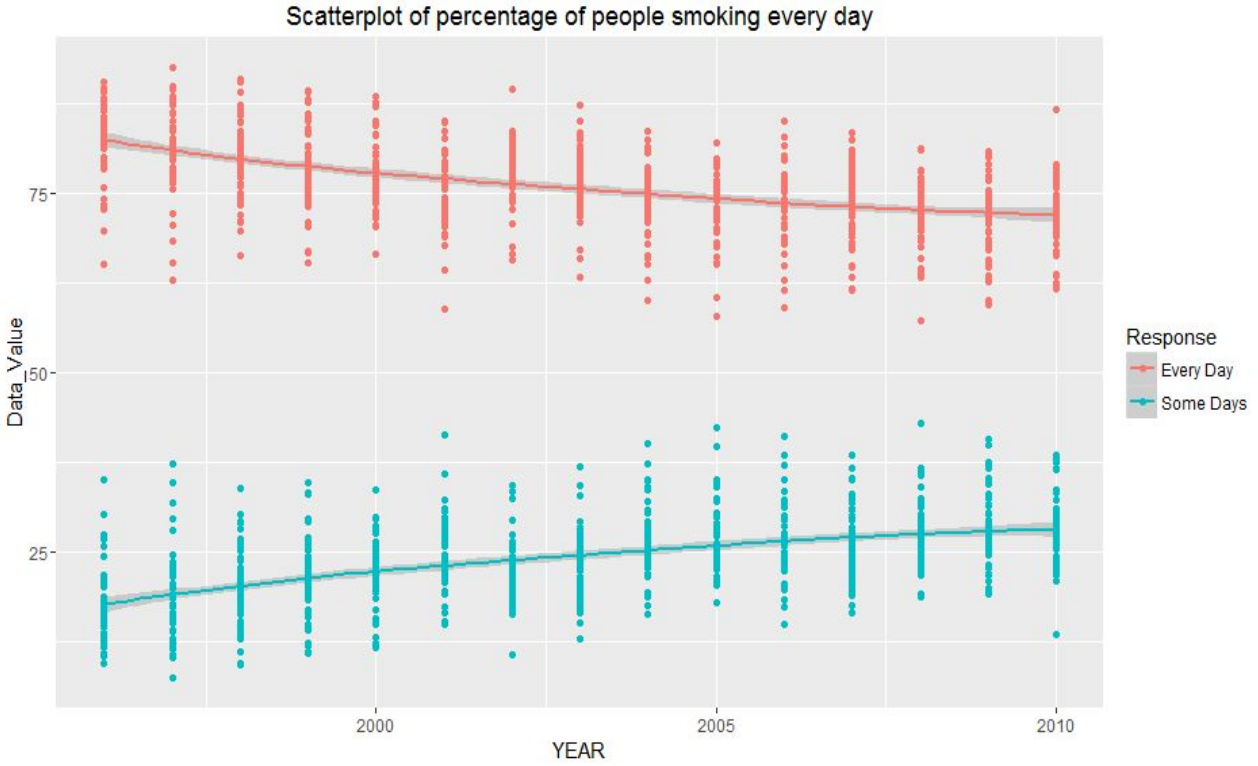
After data visualization for current smoking status, we are moving to next topic - smoking frequency.

It requires less data visualizations because it only has three variables: year, location and gender.

Smoking frequency has two responses: “Every Day” and “Some Days”.

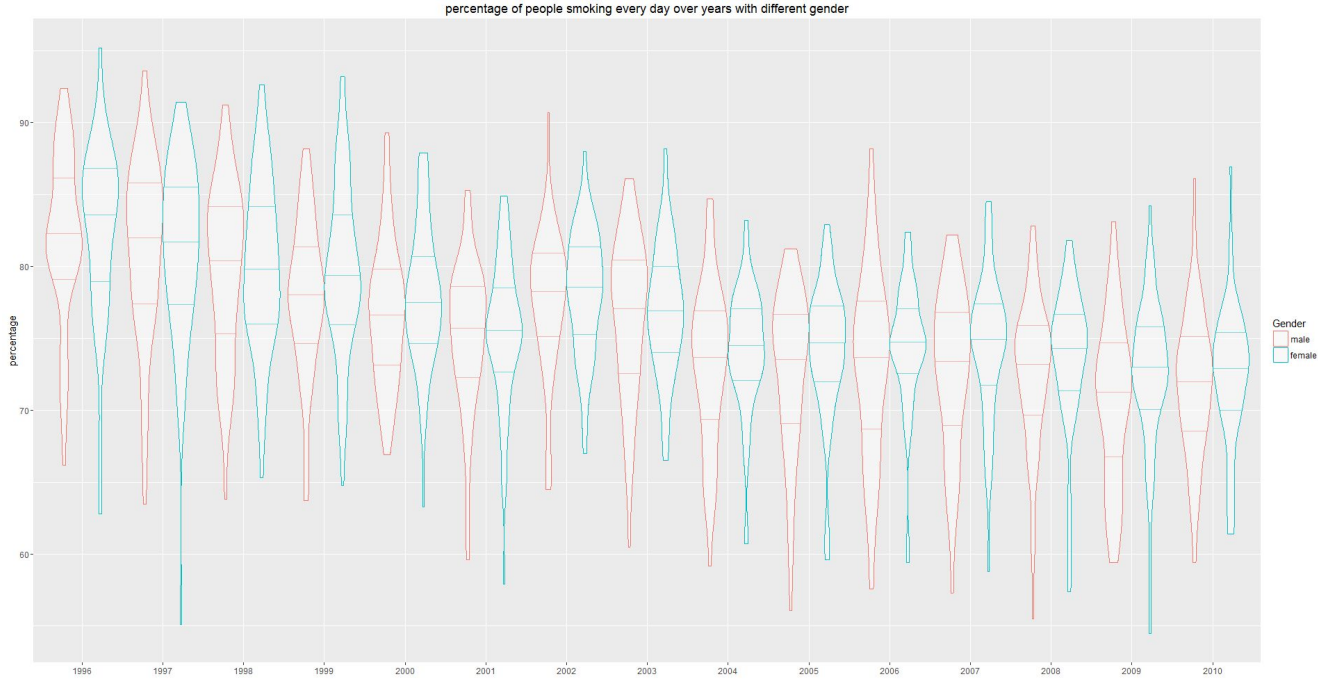
# Variable: year

The graph obviously shows the trend in percentage of every day and some day smokers among all smokers over years.



# Variable: gender

Because response have only two levels and the data values of two will always be complements, we only plot the violin plot of “Every day”.



# Smoking status

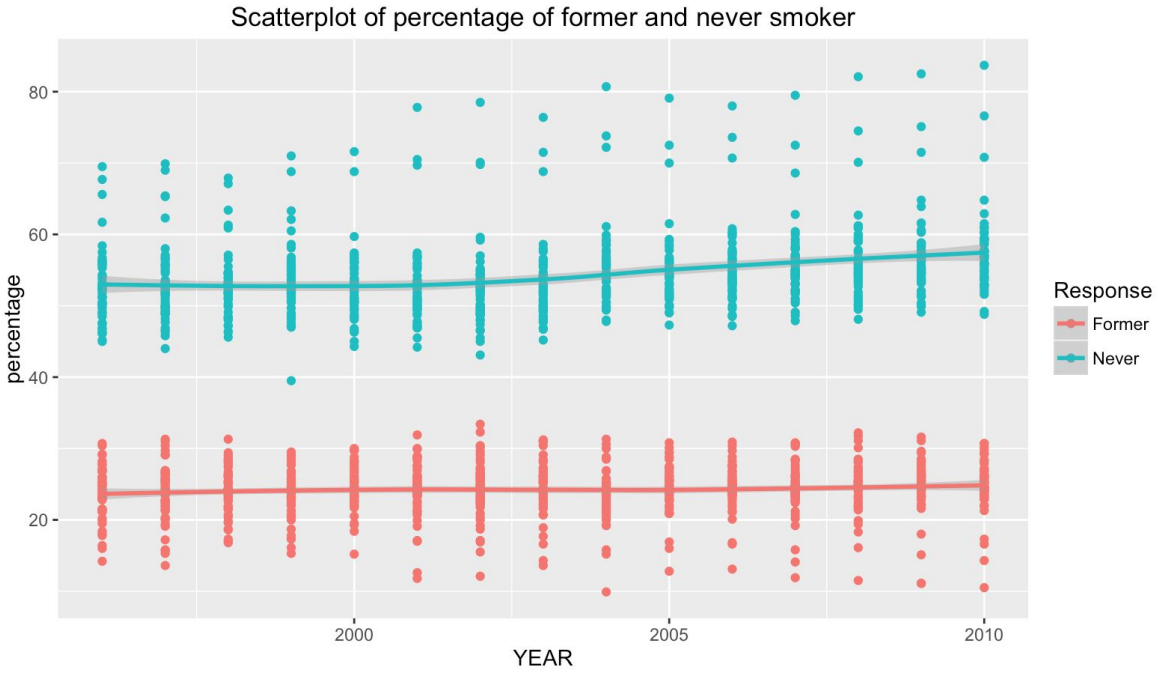
Smoking status dataset also contains only three variables: year, location and gender.

Smoking status has three responses: “current”, “never” and “former”.

we should notice that the “current” here has exactly the same values as what we have in current smoking dataset.

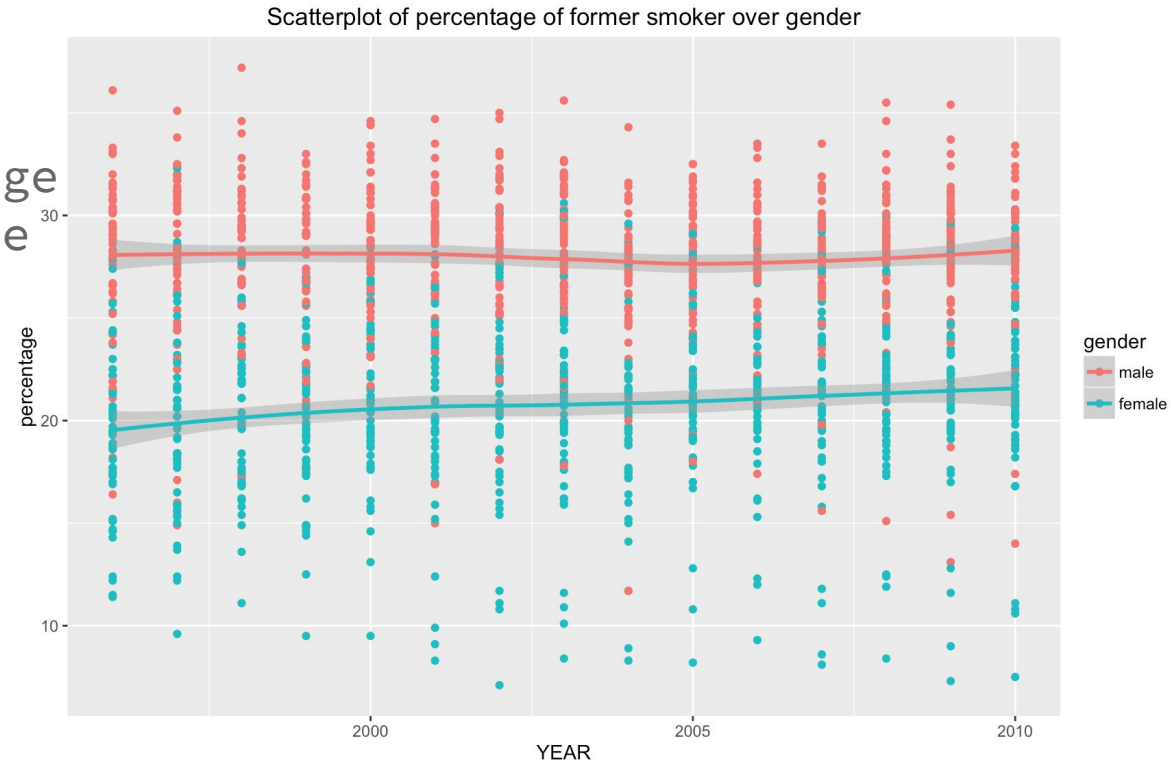
# Variable: response

distribution of people who have former smoking and who never smoke change over year.



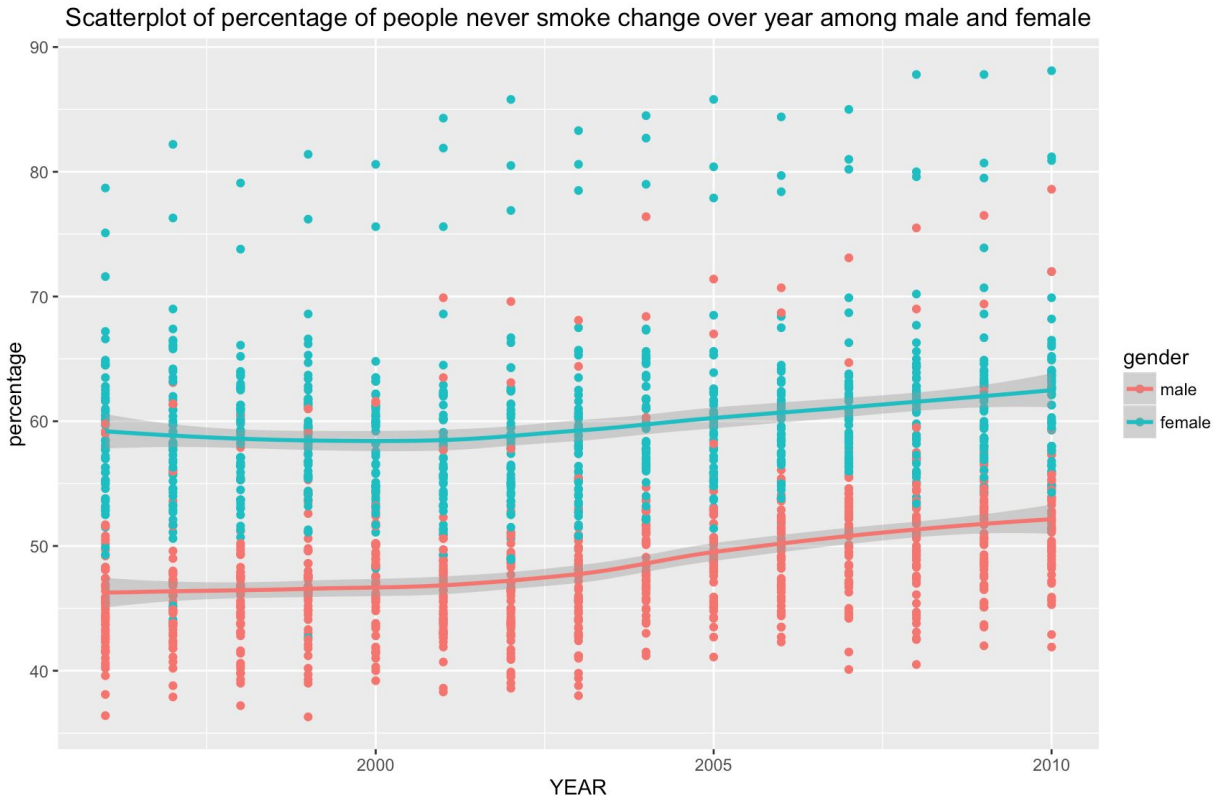
# Variable: Gender

distribution of former smokers change over year among male and female

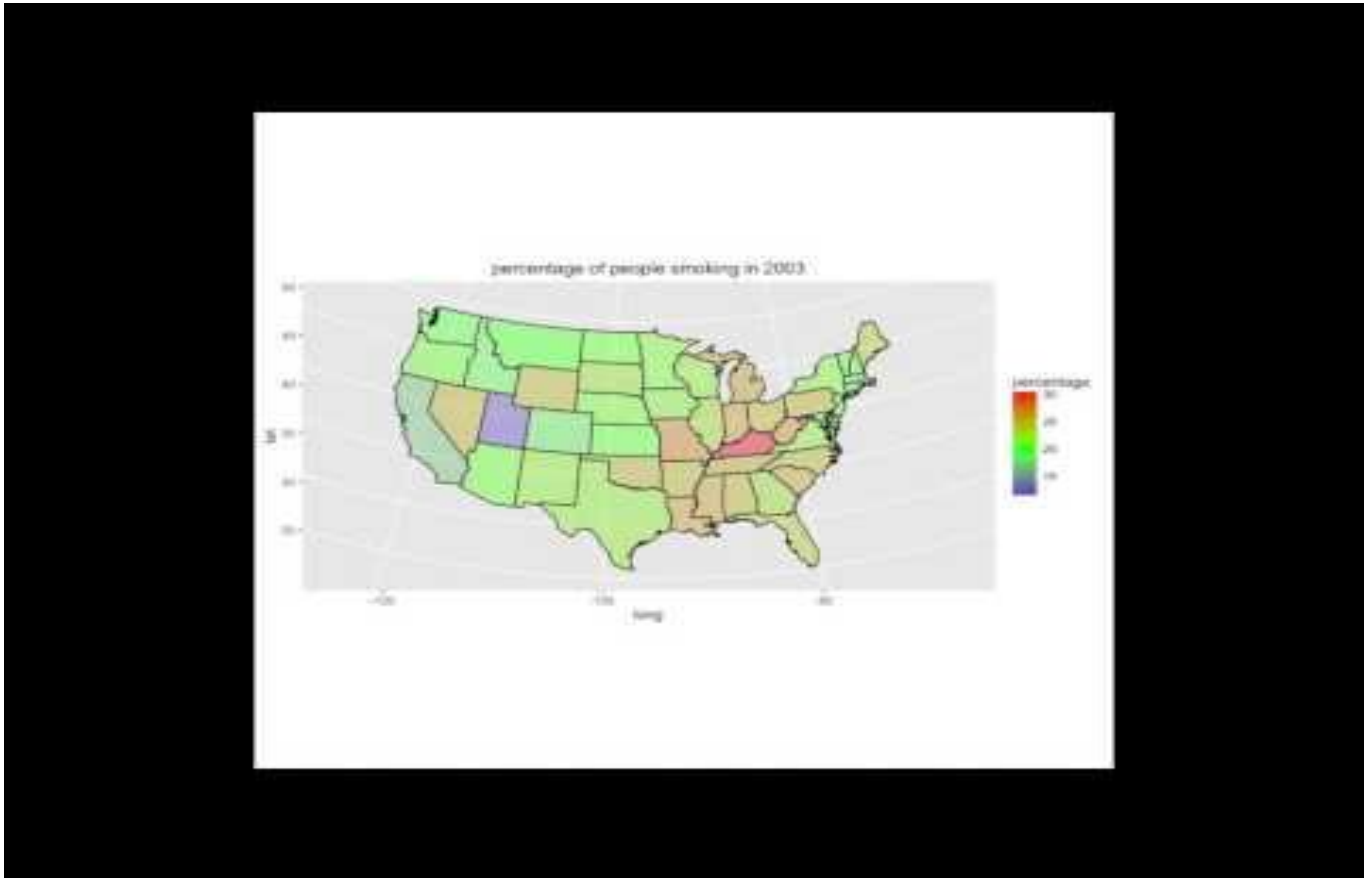


# Variable: Gender

distribution of people who never smoke change over year among male and female



# GGMAP VISUALIZATION FOR CURRENT SMOKING



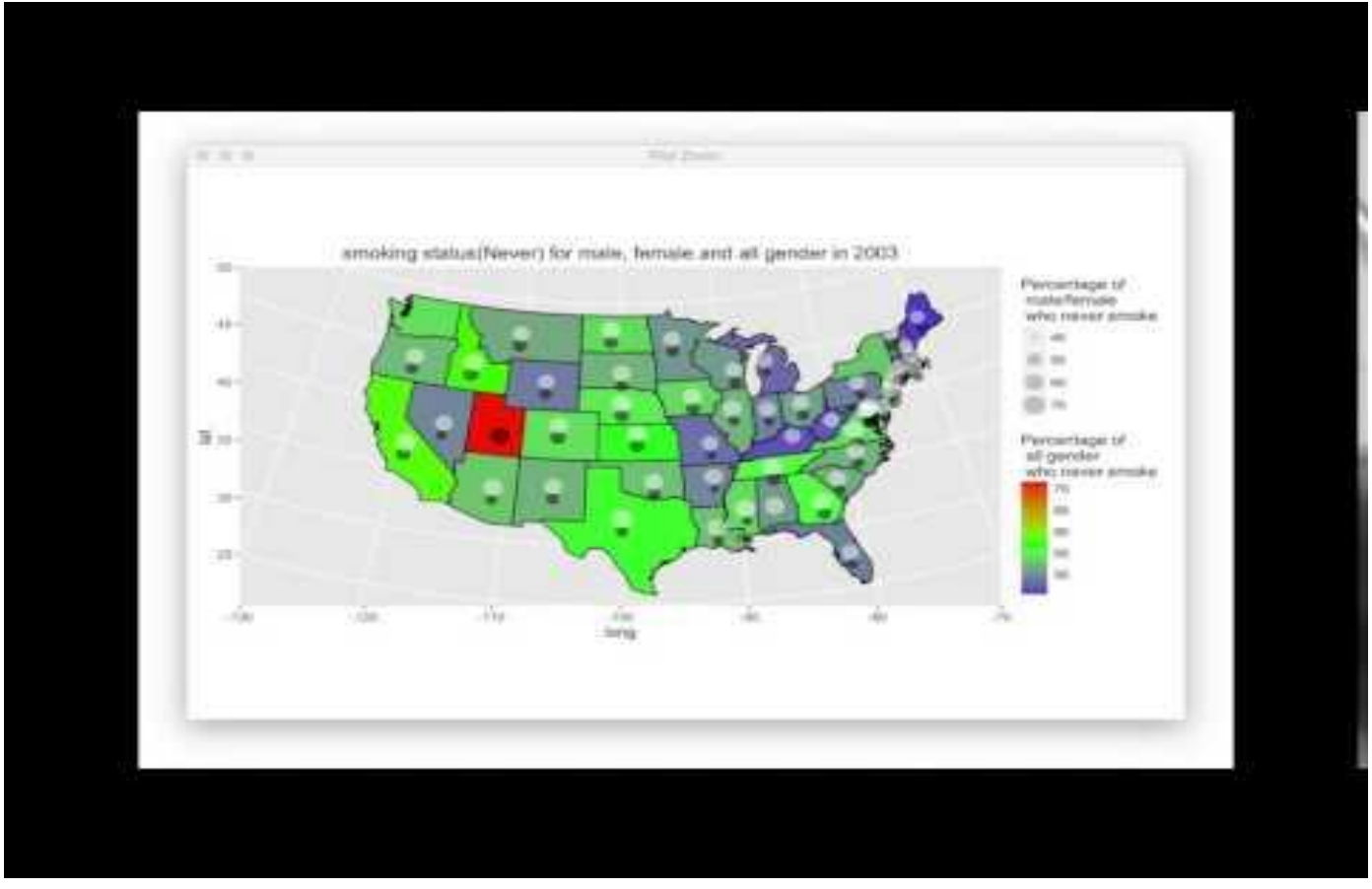
# GGMAP VISUALIZATION FOR SMOKING FREQUENCY-EVERYDAY


Black-male
white-female

# GGMAP VISUALIZATION FOR SMOKING FREQUENCY- SOME DAYS


Black-male
white-female

# GGMAP VISUALIZATION FOR SMOKING STATUS-NEVER




Black-male
white-female

# GGMAP VISUALIZATION FOR SMOKING STATUS-FORMER


Black-male
white-female