Introduction:

Question:

Methods:

Preliminary Results

Conclusion:

PM566 Final Project-Trends in the number of smokers in the United States from 1995 to 2011

Leona Ma

This is my PM566 Final Project website.

Link to **download the report** (https://raw.githubusercontent.com/LeonaMa/PM566-final-project-Trends-in-the-number-of-smokers-in-the-United-States-from-1995-to-2011/master/report.pdf)

Introduction:

Smoking can increase the risk of many diseases, such as heart disease, stroke, and lung cancer. It is one of the most important risk factors of lung cancer. Tons of efforts have been done to prevent people from smoking. This data shows that how the proportion of four-level of smoking prevalence changed in adult over 1995-2011 in US.

Question:

Was the proportion of American adults smoke decreasing during 1995-2011?

Methods:

I acquired the smoking data for 1996-2010 and 2011 separately from US Center for Disease Control and Prevention. I also extracted the longitude and latitude data from the variable Location1(state longitude latitude) in the original data (data:location (data:location)) and merge it with data from 2011 to see how the smoking status in each state of US in 2011. Since there are four levels of smoking status (Smoke everyday, Smoke some days, Former smoker, Never smoked), I also create two variables to combine the first two variables and last two variables to see the the proportion of smokers and nonsmokers overall. Then, I split the dataset into two that has state data (area) and nationwide (nation) data respectively. After all this data processing, I create several scatter plot to the the trend of tobacco use from 1995 to 2011 using ggplot and ggplotly.

```
tobacco[, smokers := `Smoke everyday`+`Smoke some days`]
tobacco[, nonsmokers := `Former smoker`+`Never smoked`]
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

Preliminary Results

Taking a look at the top5 proportions of smokers and nonsmokers.

Year	State	proportions of smokers
2003	Guam	34.1
2002	Kentucky	32.6
2002	Guam	31.9
1996	Kentucky	31.7
1999	Nevada	31.5

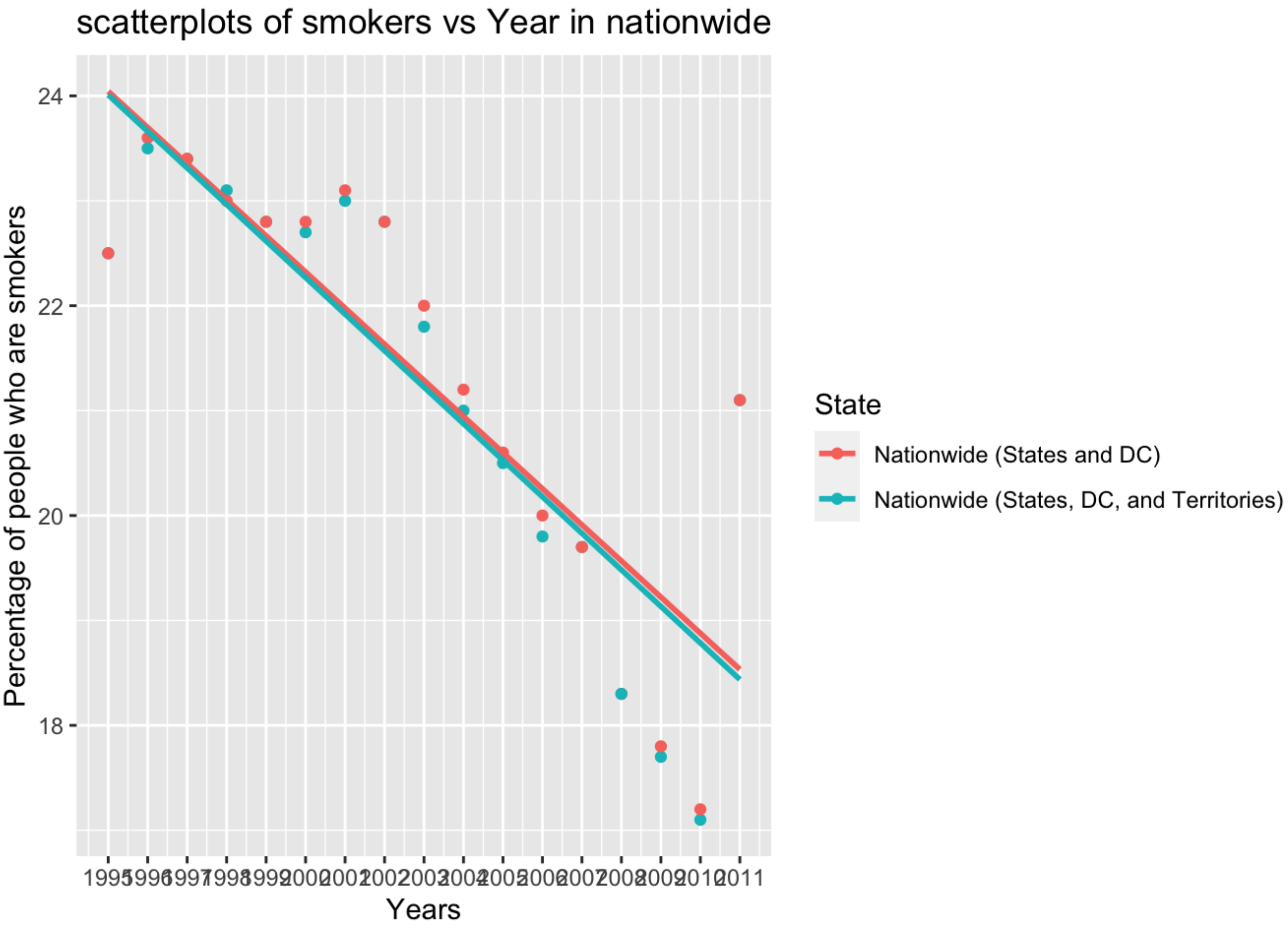
Year	State	proportions of nonsmokers
2010	Virgin Islands	94.2
2008	Virgin Islands	93.6
2009	Virgin Islands	93.6
2005	Virgin Islands	91.9
2007	Virgin Islands	91.4

From these two table, we can see that the 5 highest proportions of smokers are concentrated on 1996-2003 period, and the 5 highest proportions of nonsmokers are concentrated on 2005-2010. They show a basic trend that proportion of smokers were decreasing. However, further data exploration is still needed to make conclusion.

Looing at smokers and nonsmokers overall in nationwide and statewide respectively

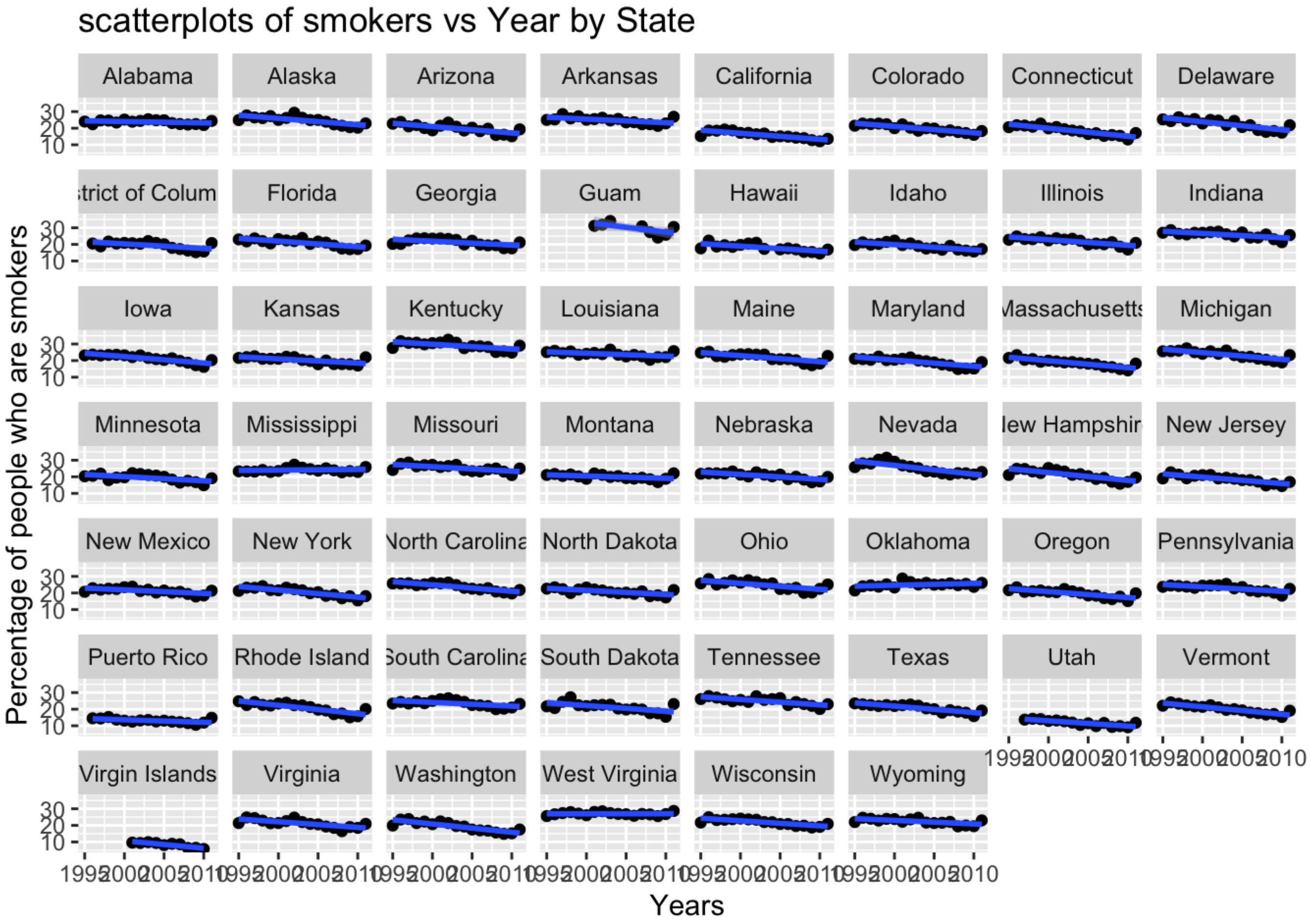
Smokers nationwide:

```
## `geom_smooth()` using formula 'y ~ x'
```


scatterplots of smokers vs Year in nationwide

From this plot we can see that adding data from territories didn’t change the trend much and it has some missing values, so just looking at the red line is fine. We can clearly see a decreasing trend in this plot. However, there is a small fluctuation between 2000-2003 and a relatively large increase from 2010 to 2011. Since including the data of territories did not change the trend much, I will not include it in my following analysis.

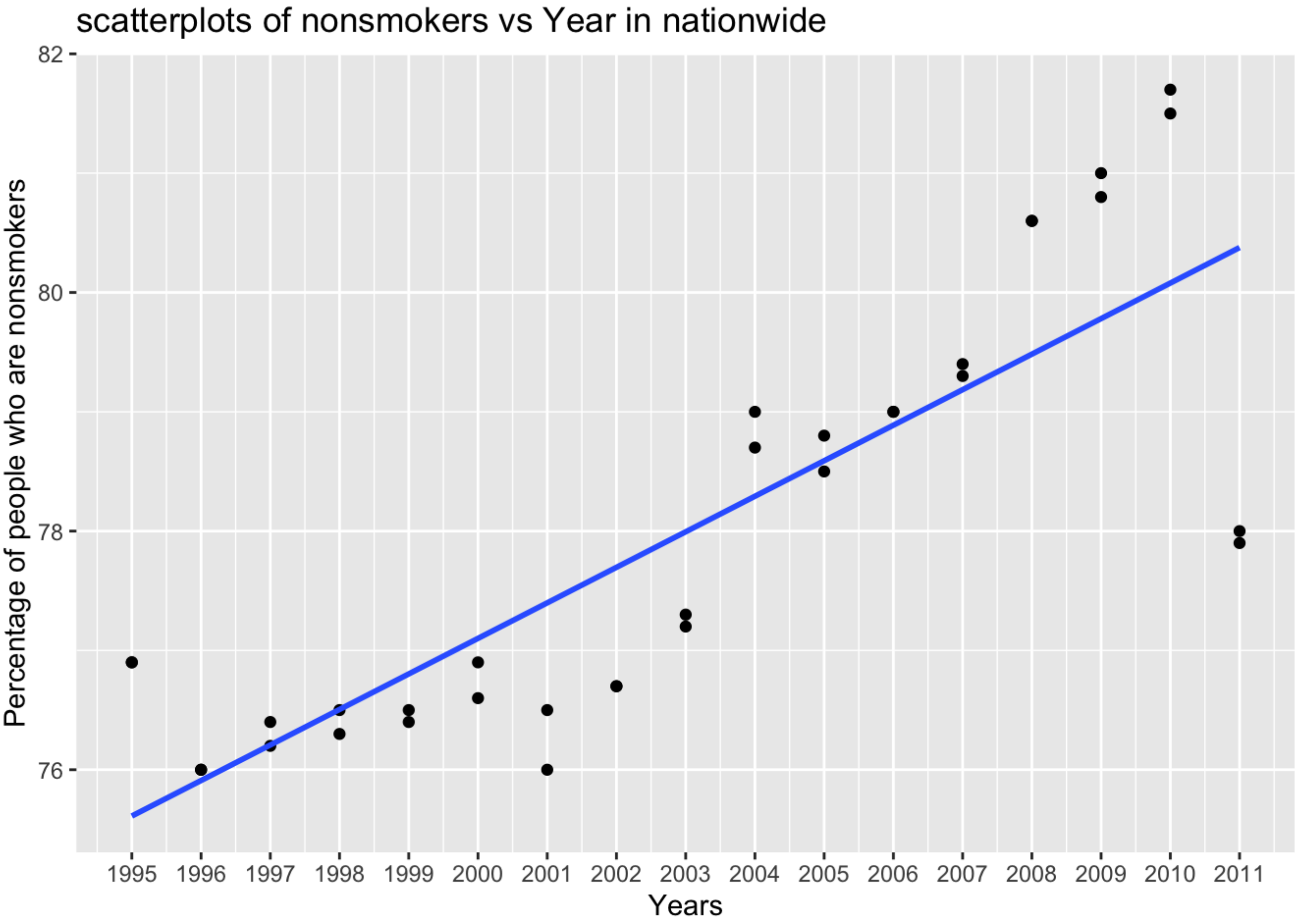
Smokers in statewide



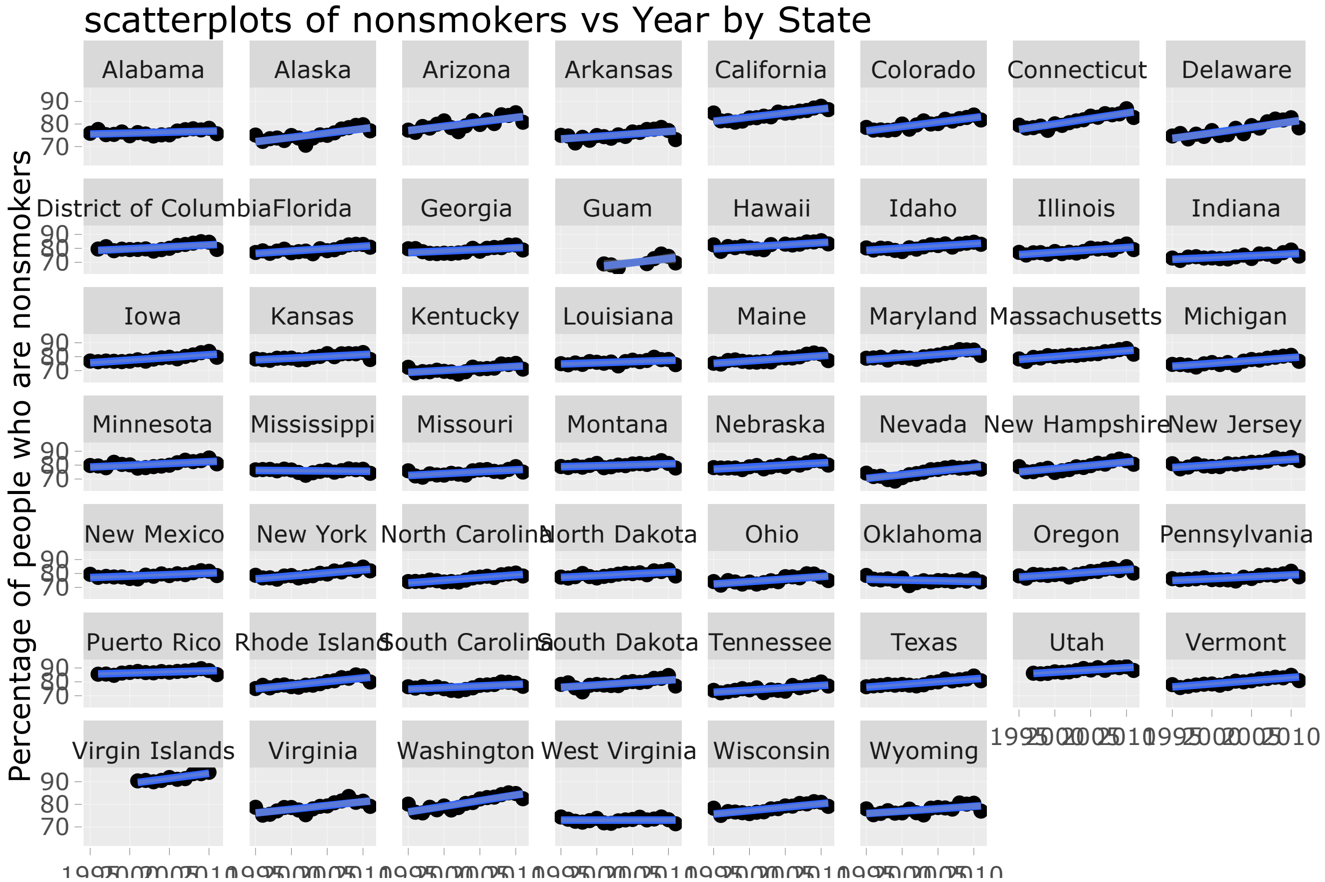
From these plots,even thought they are pretty small, we can still see that most of them have the decreasing trend of smoker proportions, and the the last point is higher than the regression line. The small fluctuation can be found in some of the area, such as South Dakota and Nevada. Thus, the result from statewide is corresponding to that of nationwide. Since all the plots has the same y-scale, we can state that the overall proportion of smokers is the highest in Kentucky, and it is relatively low in Utah and Virgin Islands.

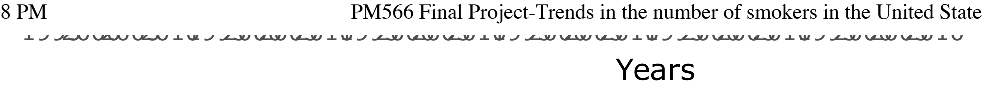
Nonsmokers nationwide:

```
## `geom_smooth()` using formula 'y ~ x'
```



Nonmokers in statewide



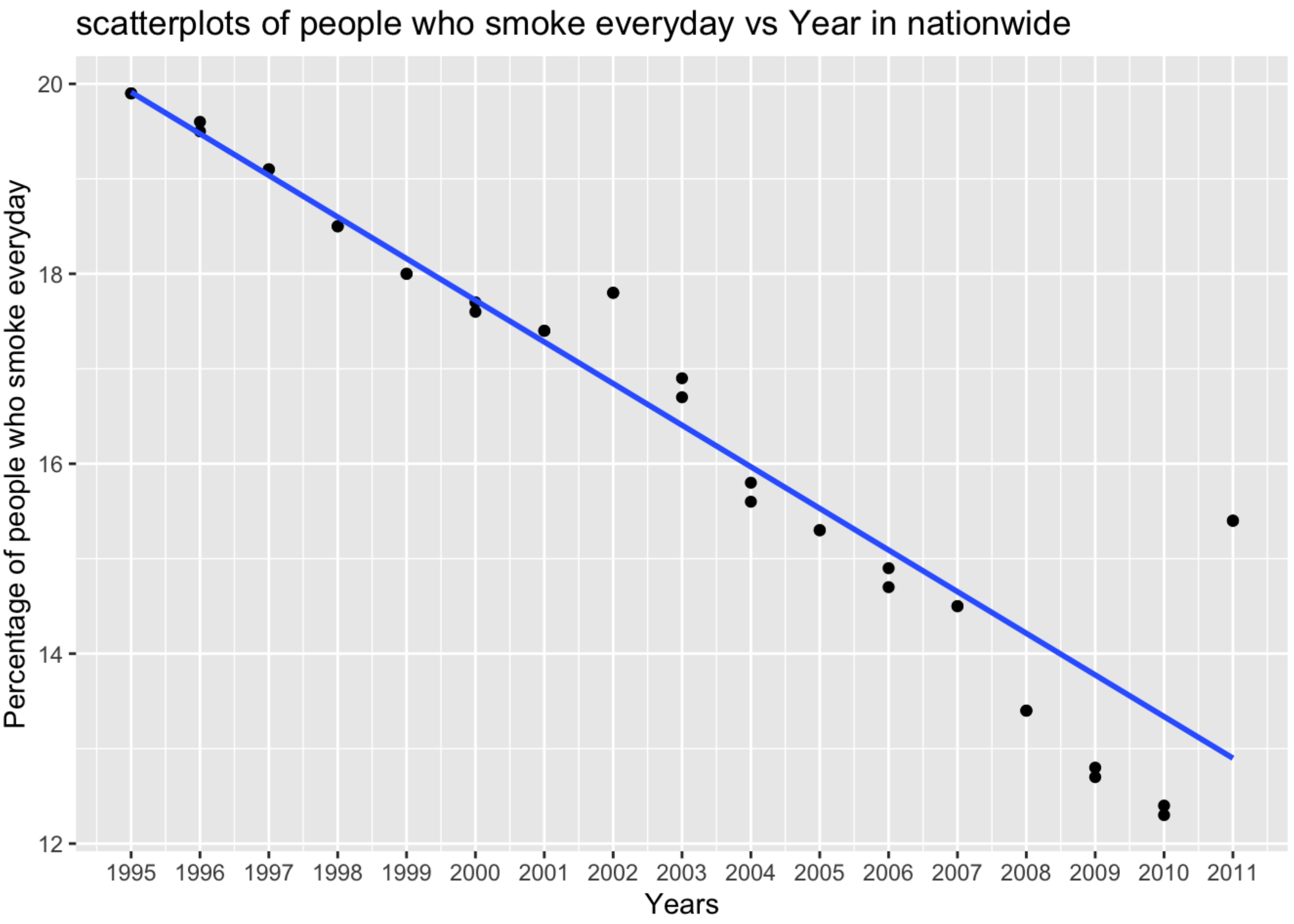


The result from proportions of nonsmokers are exactly opposite of that of the smokers. We can clearly see a increasing trend in nationwide plot.There is a small fluctuation between 2000-2003 and a relatively large decrease from 2010 to 2011. From the statewide plots, we can see that most of them have the increasing trend of nonsmoker proportions, and the the last point is lower than the regression line. The small fluctuation can be found in some of the area, such as South Dakota and Nevada. we also can state that the overall proportion of nonsmokers is the lowest in Kentucky, and it is relatively high in Utah and Virgin Islands.

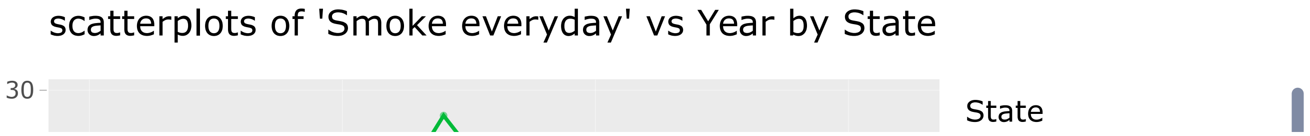
Looing at 4-level of tobacoo use in nationwide and statewide respectively

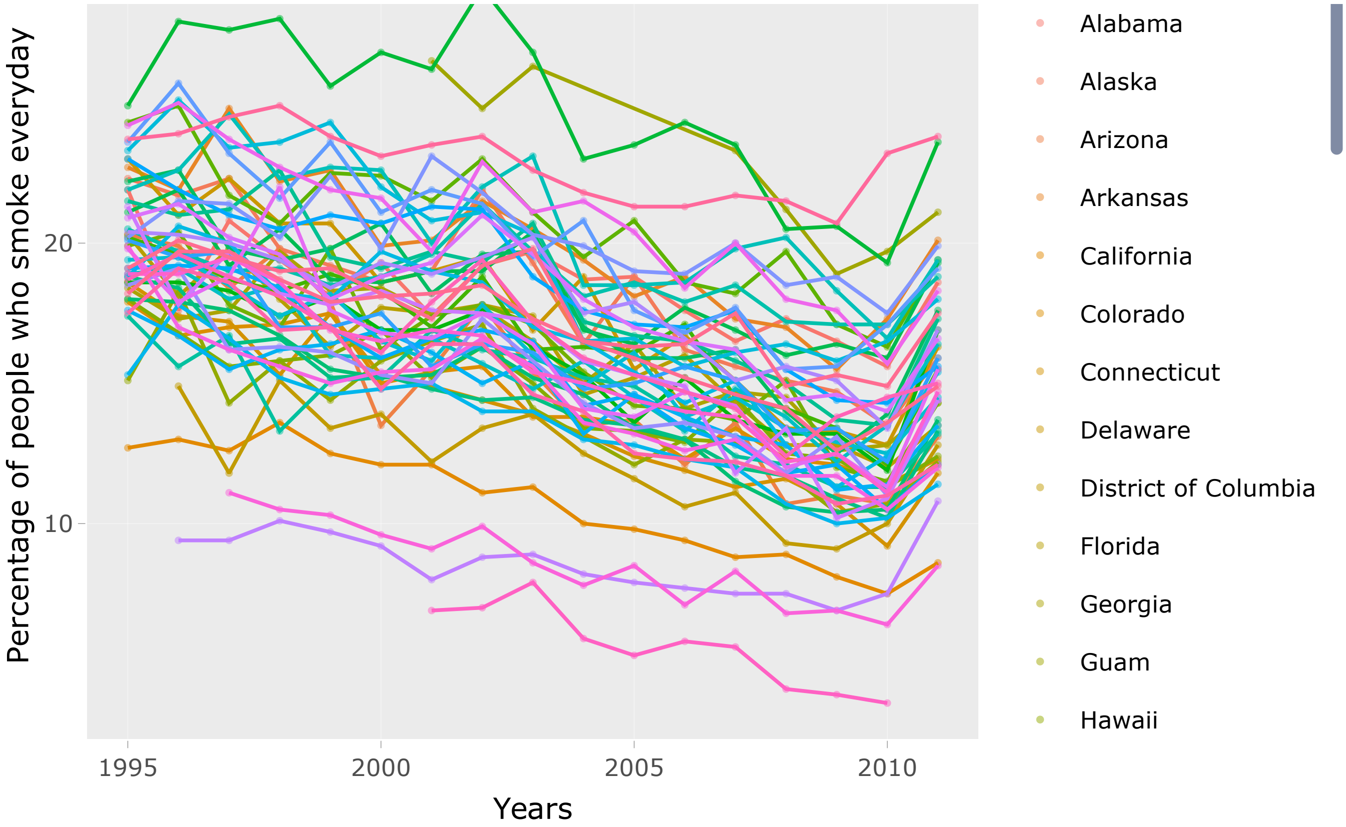
Smoke everyday in nationwide

```
## `geom_smooth()` using formula 'y ~ x'
```



Smoke everyday in statewide

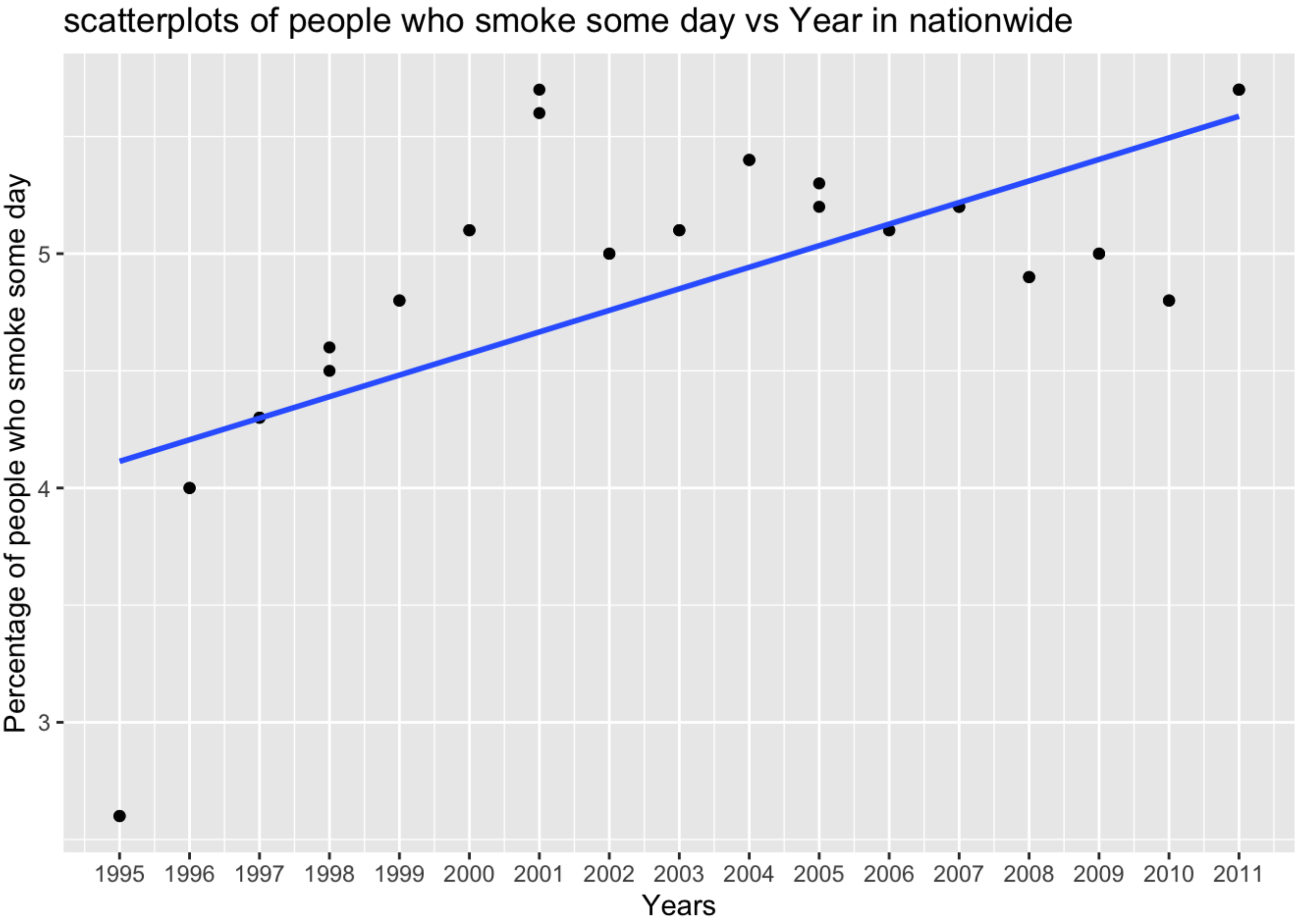


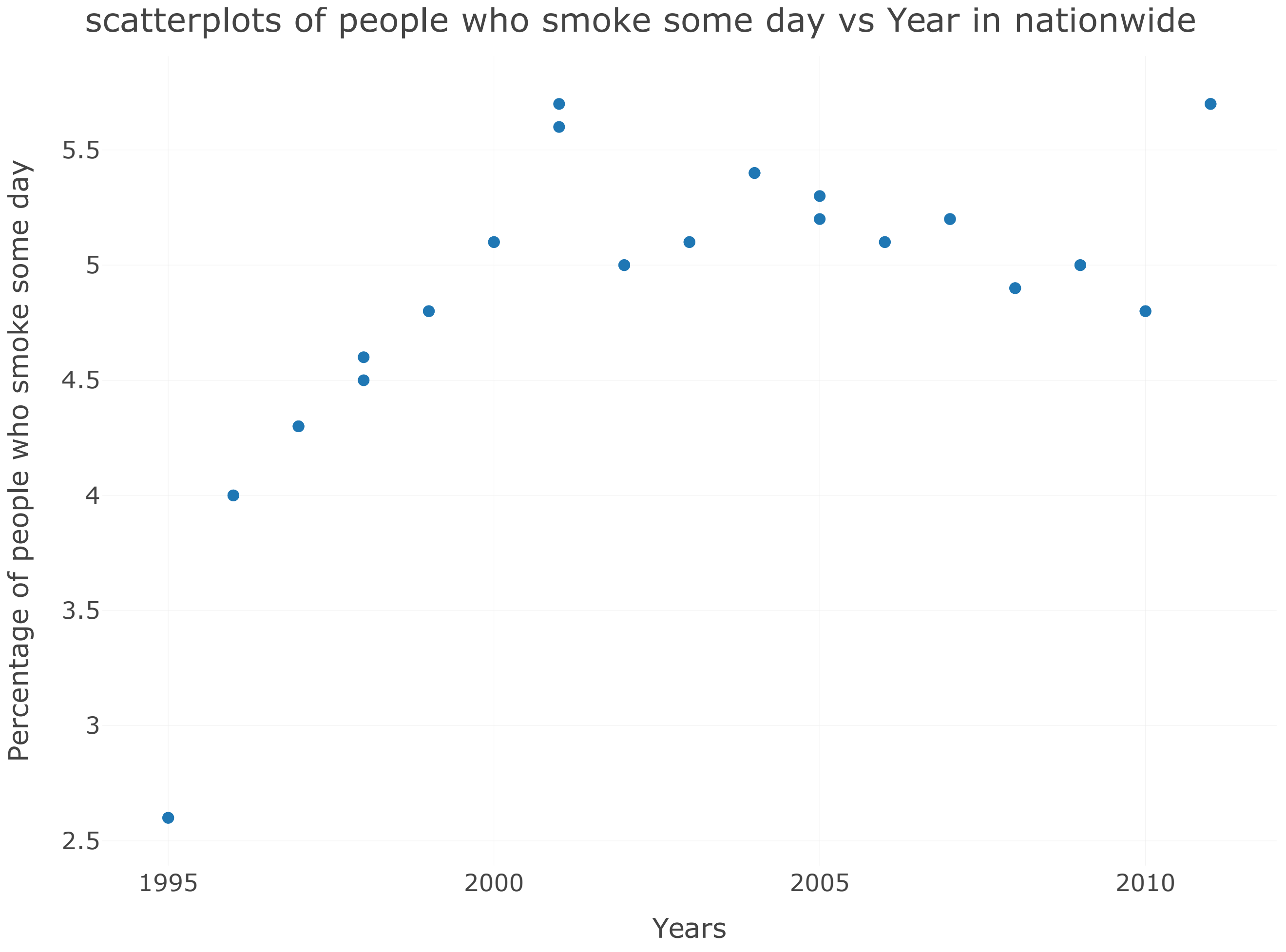


For proportions of people who smoke everyday, we can see that it is pretty similar with smokers overall, which has a clearly decreasing trend, little fluctuation and large increase in 2011. The difference is that the little fluctuation begins at 2002.

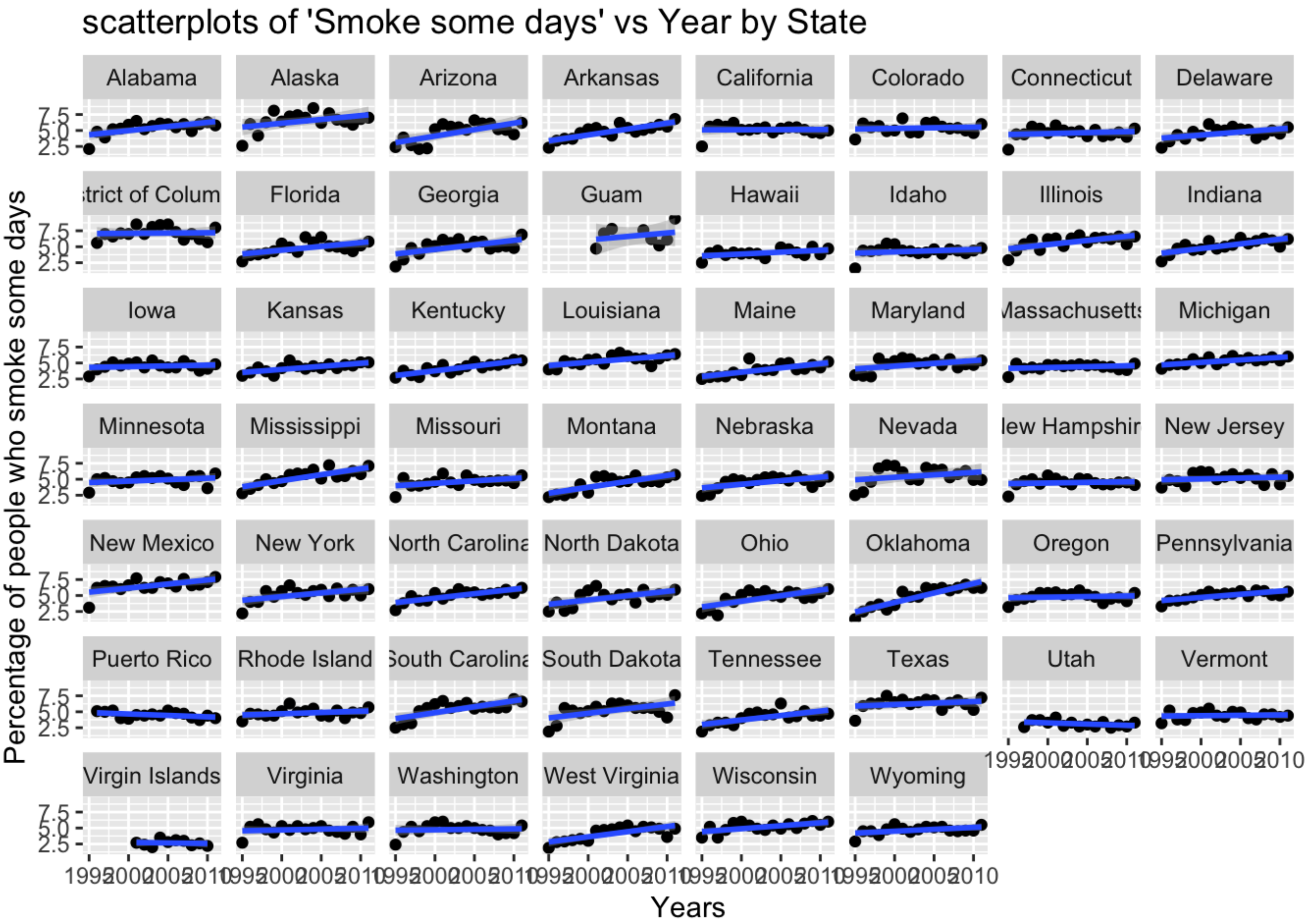
Smoke some days in nationwide

```
## `geom_smooth()` using formula 'y ~ x'
```


scatterplots of people who smoke some day vs Year in nationwide
Percentage of people who smoke some day
3
4
5
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Years


scatterplots of people who smoke some day vs Year in nationwide
Percentage of people who smoke some day
2.5
3
3.5
4
4.5
5
5.5
1995
2000
2005
2010
Years

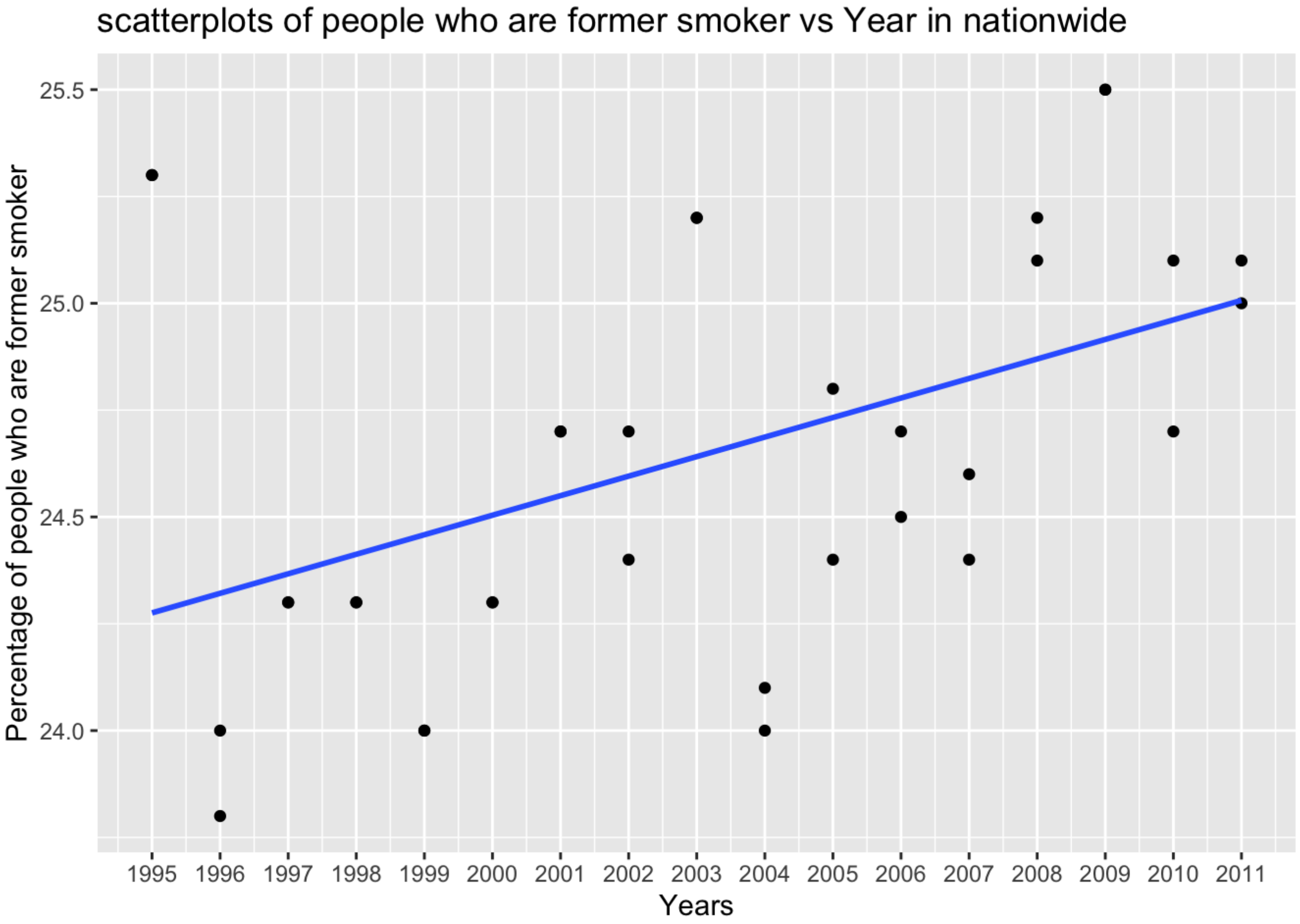
Smoke some days in statewide



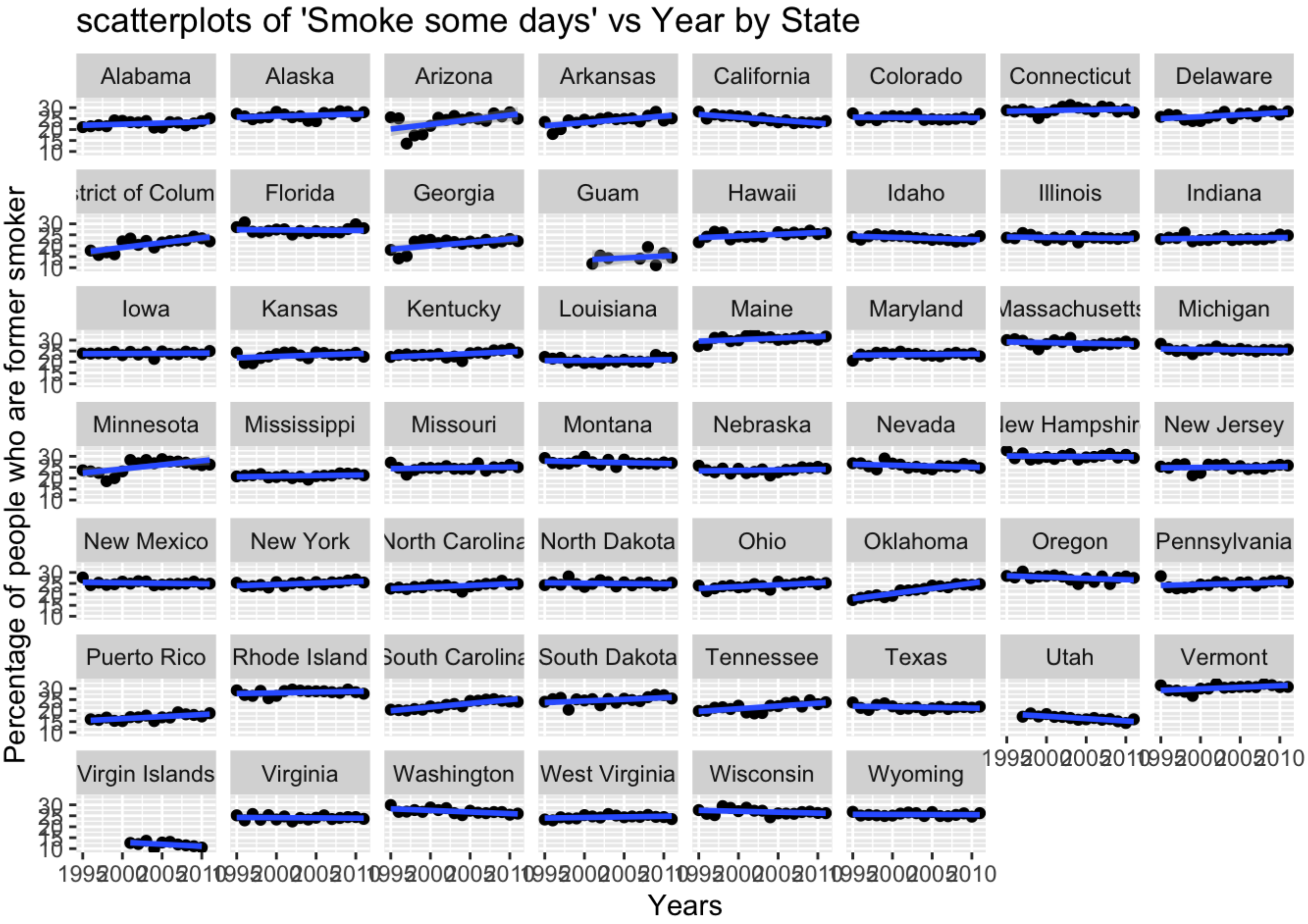
From the first nationwide plot, we can see that proportions of prople who smoke some day increased steadily from 1995 to 2001, and decreased in 2002, which corresponding to the decreasing of people who smoke everyday. We can make a reasonable inference that from 1995 to 2001, there were certain amount of people who reduced their smoking frequency, but part of them didn't stick to it, and smoked everyday again in 2002. Even though, from 2004 to 2010 there is a steady decrease, it went back agian in 2011. The statewide plots show a overall inceasing trend.

Former smoker in nationwide

```
## `geom_smooth()` using formula 'y ~ x'
```



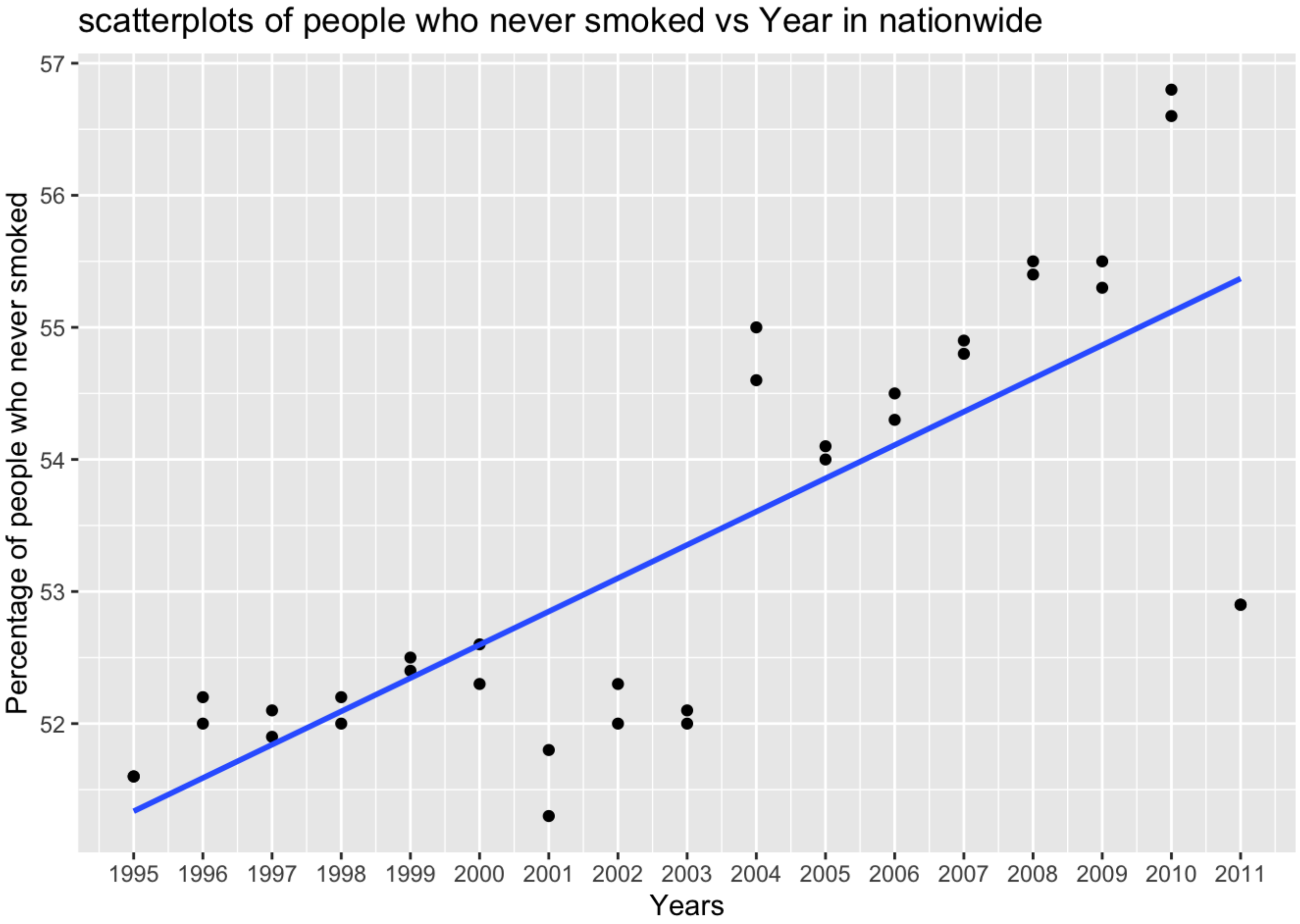
Former smoker in statewide



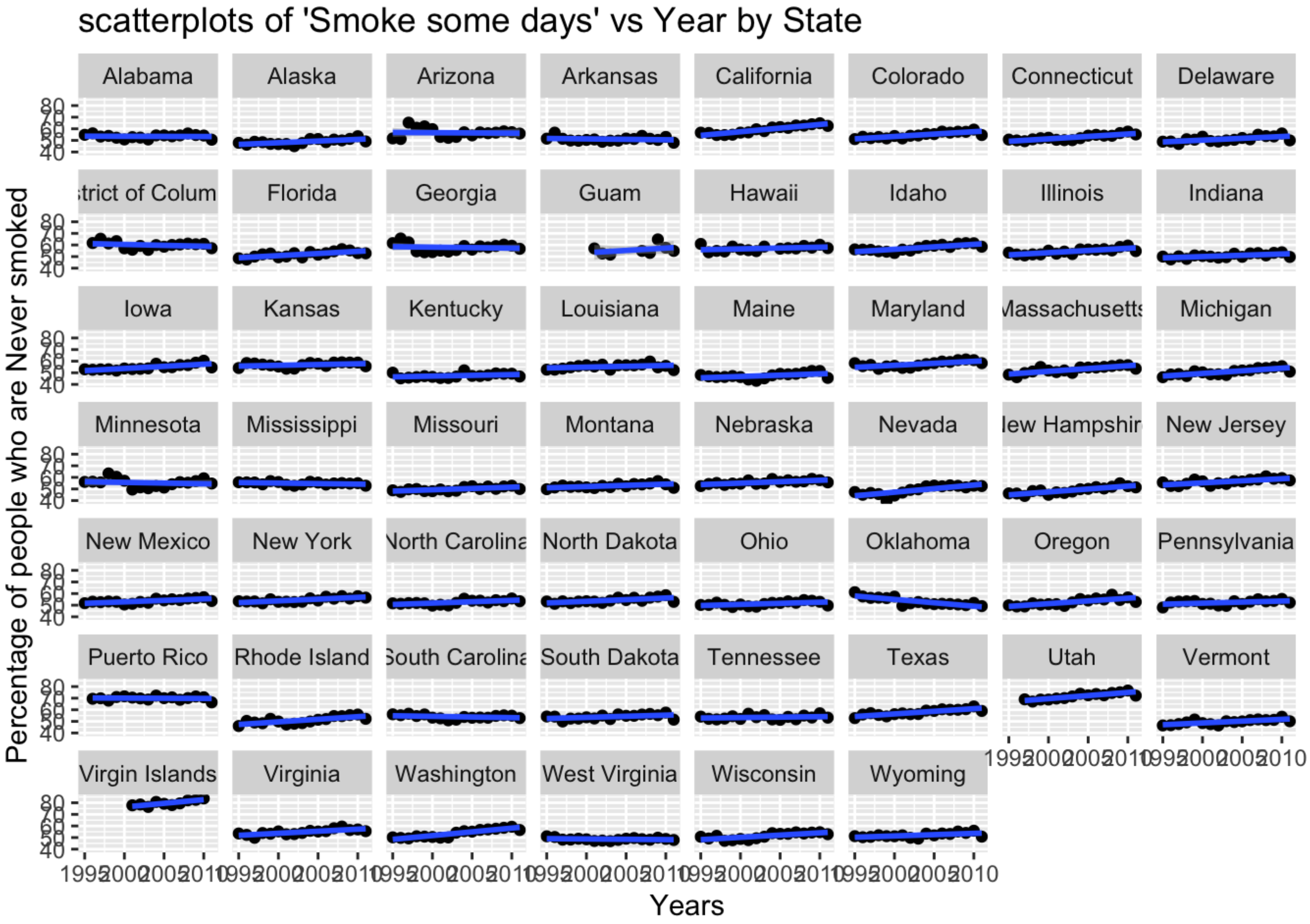
From the plot of proportions of people who are former smokers, we can see the overall trend is increasing, but it fluctuates a lot. The trend is different across each state. My inference is that there had been people try to quit smoking, but it is hard to insist.

Never smoked in nationwide

```
## `geom_smooth()` using formula 'y ~ x'
```

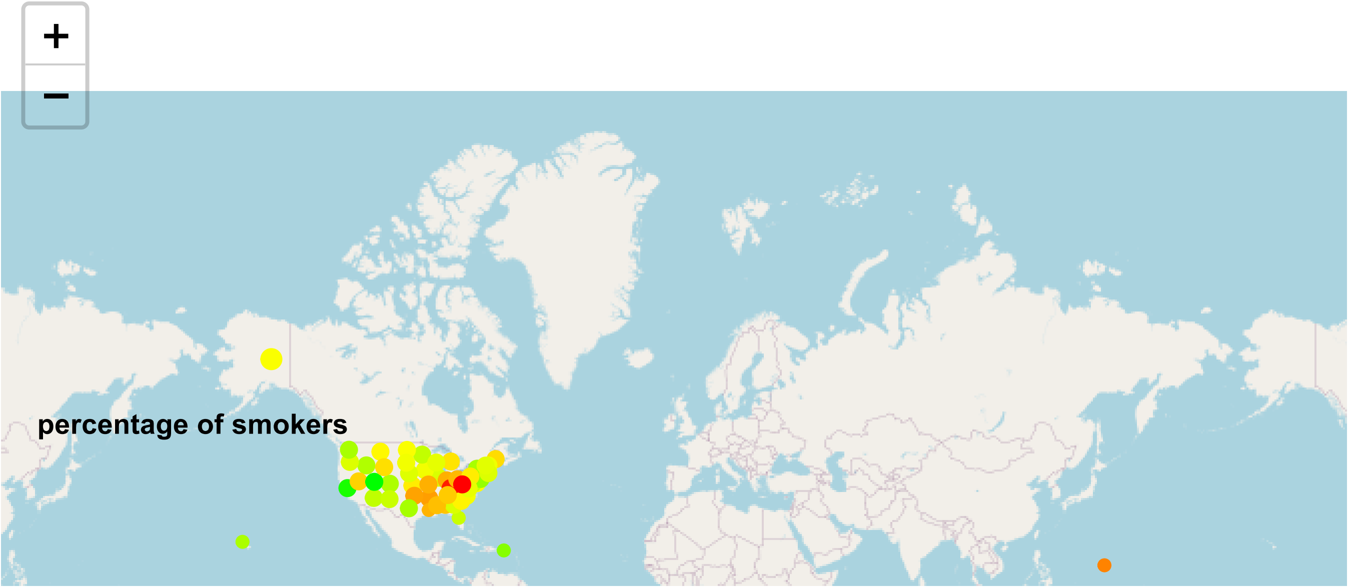


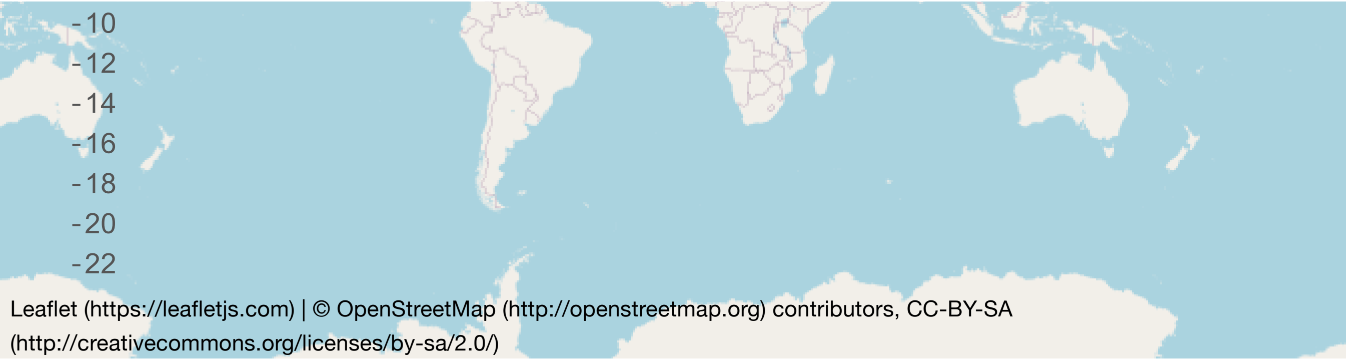
Never smoked in statewide



From the first plot of proportions of people who never smoked, we can clearly see that there is a increasing trend in the whole time period. The fluctuation between 2000 and 2003, and the large increase in 2011 are corresponding to the overall smokers and nonsmokers situation. Conditions are pretty different among each state. Most of the states have a overall increase trend or relatively steady, but some states like Oklahoma has a decreasing trend.

See percentage of people who smoke everyday over the country in most recent year, 2011





Leaflet (https://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.0/)

This plot shows that the proportion of people who smoke everyday are higher in the middle east part of America, especially in Kentucky and West Virginia, which is corresponding to what we found in our statewide plots.

Conclusion:

Generally speaking, the proportion of smokers is decreasing between 1995 and 2010 in US, but it increased a lot in the last year of observation, 2011. Thus, further data still needed since ten years have been pasted. It is highly possible that proportion of smokers has been increasing since 2011. Moreover, the reason of fluctuation between 2000-2003 is also worth exploring, since in order to reduce the number of smokers, we have to know why it increased. Even though proportion of smokers was decreasing for all the states, the exact proportions of it is quite different. Further information is still needed to figure out why.