Final Paper

Sonia Yuan & Tony Wang

Tuesday, May 03, 2016

Background

The dataset we analyzed in our final project is Behavior Risk Factor of Tobacco Use, sourcing from BRFSS survey data. The BRFSS is a continuous, state-based surveillance system that collects information about modifiable risk factors for chronic diseases and other leading causes of death. Centers for Disease Control and Prevention(CDC) is the institution that collected and arrange the data. Due to the uniqueness of the dataset, our goal is not applying validation and modelling on this dataset, but to figure out its pattern and to do visualization on it. Then, we can get a trend of people smoking in the past few years and we can make some conclusion on this trend. We would introduce how we clean and prepocess the dataset and how the data is collected in the “method” part. We will also introduce the visualizations we apply on the dataset and analysis on these visualizations in the “result” part. In the “discussion” part we will discuss our challenges in this project. The “conclusion” part would have a summary on this project. At the bottom of the paper, we will have a “citation part”, in which we would cite all the resources we have used in this final project.

Methods

The dataset records the Tobacco use of people in different states in different year. It contains 29 variables and 38050 observations.

```
head(Tobacco.FULL)
```

```
##   YEAR LocationAbbr LocationDesc                  TopicType
## 1 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
## 2 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
## 3 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
## 4 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
## 5 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
## 6 2001           AL      Alabama Tobacco Use â<U+0080><U+0093> Survey Data
##                TopicDesc     MeasureDesc DataSource Response
## 1 Cigarette Use (Adults) Current Smoking      BRFSS
## 2 Cigarette Use (Adults) Current Smoking      BRFSS
## 3 Cigarette Use (Adults) Current Smoking      BRFSS
## 4 Cigarette Use (Adults) Current Smoking      BRFSS
## 5 Cigarette Use (Adults) Current Smoking      BRFSS
## 6 Cigarette Use (Adults) Current Smoking      BRFSS
##   Data_Value_Unit Data_Value_Type Data_Value Data_Value_Footnote_Symbol
## 1               %      Percentage       23.2
## 2               %      Percentage       23.8
## 3               %      Percentage       24.6
## 4               %      Percentage       25.1
## 5               %      Percentage       25.8
## 6               %      Percentage       27.7
##   Data_Value_Footnote Data_Value_Std_Err Low_Confidence_Limit
## 1                                   1.6                 20.1
## 2                                   0.9                 22.0
## 3                                   1.1                 22.5
## 4                                   3.0                 19.3
## 5                                   1.5                 22.9
## 6                                   1.7                 24.4
##   High_Confidence_Limit Sample_Size  Gender      Race              Age
## 1                  26.3         922 Overall All Races   45 to 64 Years
## 2                  25.6        2789 Overall All Races         All Ages
## 3                  26.7        1997 Overall     White         All Ages
## 4                  30.9         253 Overall All Races   18 to 24 Years
## 5                  28.7        1048    Male All Races         All Ages
## 6                  31.0         859 Overall All Races Age 25 and Older
##    Education                              GeoLocation TopicTypeId TopicId
## 1 All Grades (32.84057112200048, -86.63186076199969)         BEH  100BEH
## 2 All Grades (32.84057112200048, -86.63186076199969)         BEH  100BEH
## 3 All Grades (32.84057112200048, -86.63186076199969)         BEH  100BEH
## 4 All Grades (32.84057112200048, -86.63186076199969)         BEH  100BEH
## 5 All Grades (32.84057112200048, -86.63186076199969)         BEH  100BEH
## 6 12th Grade (32.84057112200048, -86.63186076199969)         BEH  100BEH
##   MeasureId StratificationID1 StratificationID2 StratificationID3
## 1    110CSA              1GEN              3AGE              6RAC
## 2    110CSA              1GEN              8AGE              6RAC
## 3    110CSA              1GEN              8AGE              5RAC
## 4    110CSA              1GEN              1AGE              6RAC
## 5    110CSA              2GEN              8AGE              6RAC
## 6    110CSA              1GEN              7AGE              6RAC
##   StratificationID4
```

```
## 1              6EDU
## 2              6EDU
## 3              6EDU
## 4              6EDU
## 5              6EDU
## 6              4EDU
```

We explore each variable and figure out that there are 3 topics “Cessation (Adults)”, “Cigarette Consumption (Adults)” and “Cigarette Use (Adults)”. Since the dataset is too huge, we decide to explore only one topic, which is “Cigarette Use (Adults)” and thus we deleted the observations belonged to the other two topics. Then, we only have 28972 observations left.

```
levels(Tobacco$TopicDesc)
```

```
## [1] "Cessation (Adults)"            "Cigarette Consumption (Adults)"
## [3] "Cigarette Use (Adults)"
```

```
Tobacco.CU<-Tobacco[Tobacco[,5]!="Cessation (Adults)" &  Tobacco[,5]!="Cigarette Consumpt
ion (Adults)",]
```

Also, we find out that some variables measures the same thing and some variables do not contain any value. Therefore, we clean up these variables. Now, we have only 11 variables in which we have 3 numerical variables and 8 categorical variables (we consider YEAR as a categorical variable).We also learn that each observation, instead of representing one person, actually represents a group of people that have the same characteristics. We can simply tell that by understanding each variable: By exploring the dataset, we know that “YEAR” shows that the observation is made in certain year; “LocationAbbr” measures where the observation was made; “Responce” indicate that there is at least one question being asked and the “Responce” variable measures the responces for teh question; “Data_Value” measures the percentage of people fulfill the characteristics, which does not include the “Responce”, described in the row has such responce to the question; “SampleSize” measures the number of people fulfill the characteristics, which does not include the “Responce”, described in the row; “GeoLocation” is the center of the state; “MeasureId” represents the question being asked for each row; “StratificationID1”, “StratificationID2”, “StratificationID3”, “StratificationID4” measure gender, age, race and education level.In order to better understand the dataset, We explore the observations happened in Alabama in 2000 and we set them to be all gender, all age, all race and all education level. Then we find out that there are three questions being asked. Also, if we add up the percentage of the rows that are asked the same question, the total percentage would be 100 percent. Also, the observation with “current” responce and “165SSA” has the same percentage value with the percentage of the observation with “110CSA”. Since “MeasureDesc” measures the same thing with the variable “MeasureID”, we know that “110CSA” asks if the person in certain group currently smoke, “166SSP” asks smoking frequency that has two responces “everyday” and “some days and”165SSA" asks smoking status that has three responces “current”, “former” and “never”.

```
Tobacco.CU1<-Tobacco.CU[,-c(3,4,5,6,7,9,10,12,13,14,15,16,18,19,20,21,23,24)]
Tobacco.CU1[Tobacco.CU1[,1]==2000 & Tobacco.CU1[,9]=="8AGE" & Tobacco.CU1[,8]=="1GEN" & T
obacco.CU1[,10]=="6RAC" & Tobacco.CU1[,11]=="6EDU" &  Tobacco.CU1[,2]=="AL" , ]
```

```
##     YEAR LocationAbbr  Response Data_Value Sample_Size
## 501 2000           AL                 25.2        2234
## 515 2000           AL Every Day       76.4         549
## 519 2000           AL Some Days       23.6         549
## 521 2000           AL   Current       25.2        2234
## 524 2000           AL    Former       24.0        2234
## 527 2000           AL     Never       50.8        2234
##                                GeoLocation MeasureId StratificationID1
## 501 (32.84057112200048, -86.63186076199969)    110CSA              1GEN
## 515 (32.84057112200048, -86.63186076199969)    166SSP              1GEN
## 519 (32.84057112200048, -86.63186076199969)    166SSP              1GEN
## 521 (32.84057112200048, -86.63186076199969)    165SSA              1GEN
## 524 (32.84057112200048, -86.63186076199969)    165SSA              1GEN
## 527 (32.84057112200048, -86.63186076199969)    165SSA              1GEN
##     StratificationID2 StratificationID3 StratificationID4
## 501              8AGE              6RAC              6EDU
## 515              8AGE              6RAC              6EDU
## 519              8AGE              6RAC              6EDU
## 521              8AGE              6RAC              6EDU
## 524              8AGE              6RAC              6EDU
## 527              8AGE              6RAC              6EDU
```

Once we know that the dataset contains only three questions and “Data_Value” measures the percentage of people in centain group has certain answer to the question, we can seperate the dataset into three datasets and we would have a dataset for the current smoking, one for smoking frequency and one for smoking status. By exploring each dataset, We found that in smoking frequency dataset only gender variable has different values whearas the other variables are all set to be “all ages”,“all education level” and “all races”, this also happend for smoking status dataset. Then we delete the variables that have the same value for all observations. We also figure out that in current smoking dataset, if the value of Age changes, the value of gender, education level and race would not change, so do race and gender. However, when education level changes, age would also change but it only change between two values “20 and above” and “25 and above”.

```
# current smoking
Tobacco.CU1c<-Tobacco.CU1[Tobacco.CU1[,7]=="110CSA",]
Tobacco.CU1c<-Tobacco.CU1c[ ,-7]
# current smoking changes over race, "6RAC" means "all races"
Tran3.year<- Tobacco.CU1c[Tobacco.CU1c[,7]=="1GEN" & Tobacco.CU1c[,8]=="8AGE" & Tobacco.C
U1c[,10]=="6EDU" & Tobacco.CU1c[,9]!= "6RAC", ]
nrow(Tran3.year[Tran3.year[,7]!="1GEN"|Tran3.year[,8]!="8AGE"|Tran3.year[,10]!="6EDU",])
```

```
## [1] 0
```

```
#changes over gender, "1GEN" means "overal"
Tobacco.CU1c[Tobacco.CU1c[,1]==2000 & Tobacco.CU1c[,2]=="AL" & c(Tobacco.CU1c[,7]=="2GEN"
| Tobacco.CU1c[,7]=="3GEN"), ]
```

```
##     YEAR LocationAbbr Response Data_Value Sample_Size
## 496 2000           AL                 22.0        1398
## 504 2000           AL                 28.9         836
##                                GeoLocation StratificationID1
## 496 (32.84057112200048, -86.63186076199969)              3GEN
## 504 (32.84057112200048, -86.63186076199969)              2GEN
##     StratificationID2 StratificationID3 StratificationID4
## 496              8AGE              6RAC              6EDU
## 504              8AGE              6RAC              6EDU
```

```
# current smoking changes over ages
levels(Tobacco.CU[,"Age"])
```

```
## [1] "18 to 24 Years"     "18 to 44 Years"     "25 to 44 Years"
## [4] "45 to 64 Years"     "65 Years and Older" "Age 20 and Older"
## [7] "Age 25 and Older"   "All Ages"
```

```
levels(Tobacco.CU1c[,"StratificationID3"])
```

```
## [1] "1RAC" "2RAC" "3RAC" "4RAC" "5RAC" "6RAC"
```

```
Tobacco.CU1c[Tobacco.CU1c[,1]==2000 & Tobacco.CU1c[,2]=="AL" & c(Tobacco.CU1c[,8]=="1AGE"
| Tobacco.CU1c[,8]=="2AGE" | Tobacco.CU1c[,8]=="3AGE" | Tobacco.CU1c[,8]=="4AGE") , ]
```

```
##     YEAR LocationAbbr Response Data_Value Sample_Size
## 493 2000           AL                 11.1         431
## 498 2000           AL                 24.0         750
## 505 2000           AL                 30.5         839
## 506 2000           AL                 32.3         200
##                                GeoLocation StratificationID1
## 493 (32.84057112200048, -86.63186076199969)              1GEN
## 498 (32.84057112200048, -86.63186076199969)              1GEN
## 505 (32.84057112200048, -86.63186076199969)              1GEN
## 506 (32.84057112200048, -86.63186076199969)              1GEN
##     StratificationID2 StratificationID3 StratificationID4
## 493              4AGE              6RAC              6EDU
## 498              3AGE              6RAC              6EDU
## 505              2AGE              6RAC              6EDU
## 506              1AGE              6RAC              6EDU
```

```
# current smoking changes over education level
levels(Tobacco.CU[,"Education"])
```

```
## [1] "< 12th Grade" "> 12th Grade" "12th Grade"   "All Grades"
```

```
levels(Tobacco.CU1c[,"StratificationID4"])
```

```
## [1] "3EDU" "4EDU" "5EDU" "6EDU"
```

```
Tobacco.CU1c[Tobacco.CU1c[,1]==2000 & Tobacco.CU1c[,2]=="AL" & c(Tobacco.CU1c[,10]=="3ED
U" | Tobacco.CU1c[,10]=="4EDU" | Tobacco.CU1c[,10]=="5EDU"), ]
```

```
##     YEAR LocationAbbr Response Data_Value Sample_Size
## 495 2000           AL                20.2         974
## 497 2000           AL                22.0        1060
## 499 2000           AL                24.5         731
## 500 2000           AL                24.8         680
## 507 2000           AL                33.4         363
## 508 2000           AL                33.9         378
##                               GeoLocation StratificationID1
## 495 (32.84057112200048, -86.63186076199969)              1GEN
## 497 (32.84057112200048, -86.63186076199969)              1GEN
## 499 (32.84057112200048, -86.63186076199969)              1GEN
## 500 (32.84057112200048, -86.63186076199969)              1GEN
## 507 (32.84057112200048, -86.63186076199969)              1GEN
## 508 (32.84057112200048, -86.63186076199969)              1GEN
##     StratificationID2 StratificationID3 StratificationID4
## 495              7AGE              6RAC              5EDU
## 497              6AGE              6RAC              5EDU
## 499              6AGE              6RAC              4EDU
## 500              7AGE              6RAC              4EDU
## 507              7AGE              6RAC              3EDU
## 508              6AGE              6RAC              3EDU
```

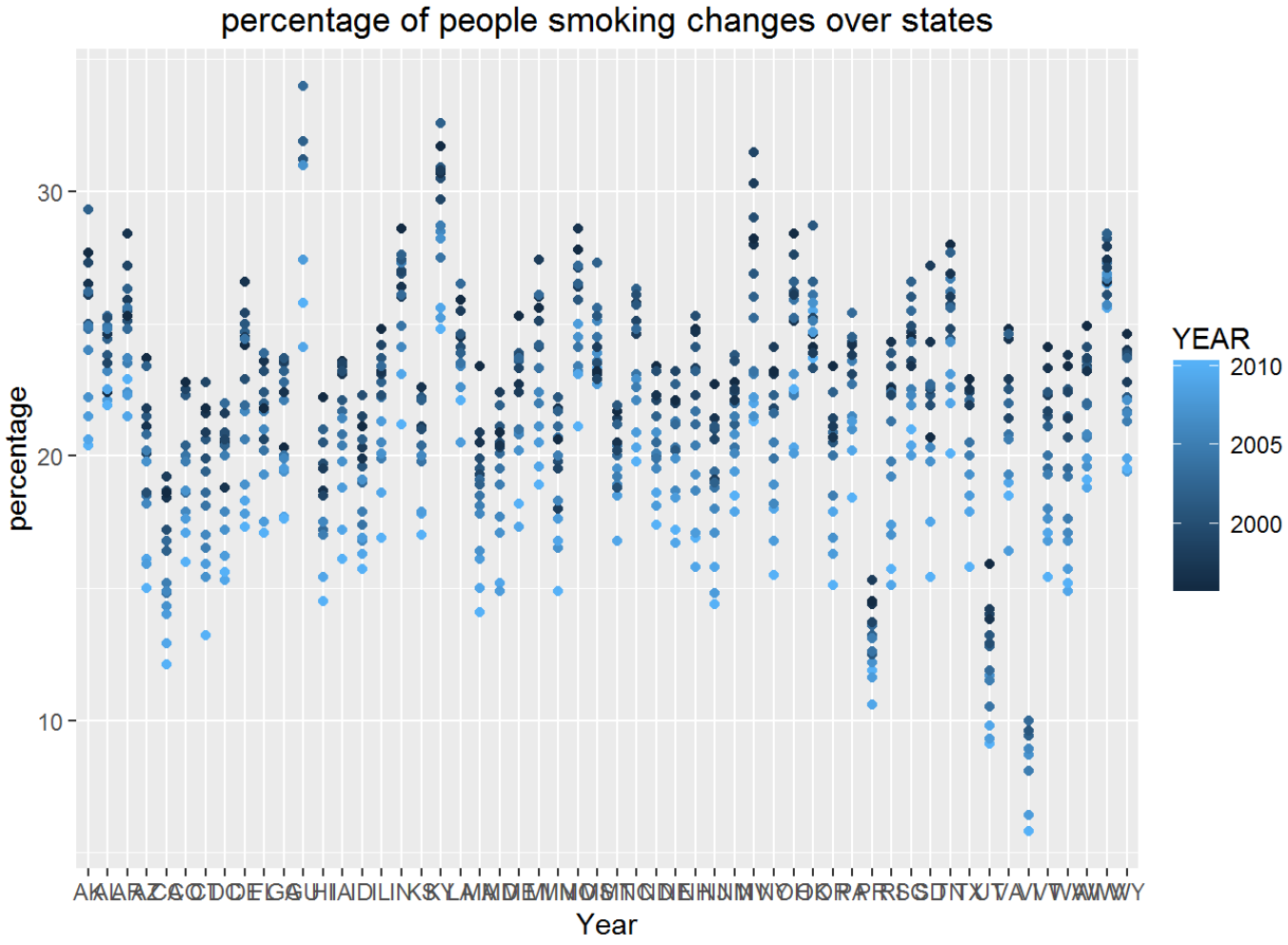
```
# smoking frequency
Tobacco.CU1f<-Tobacco.CU1[Tobacco.CU1[,7]=="166SSP",]
nrow(Tobacco.CU1f[Tobacco.CU1f[,9]!="8AGE" | Tobacco.CU1f[,10]!="6RAC"| Tobacco.CU1f[,1
1]!="6EDU", ])
```

```
## [1] 0
```

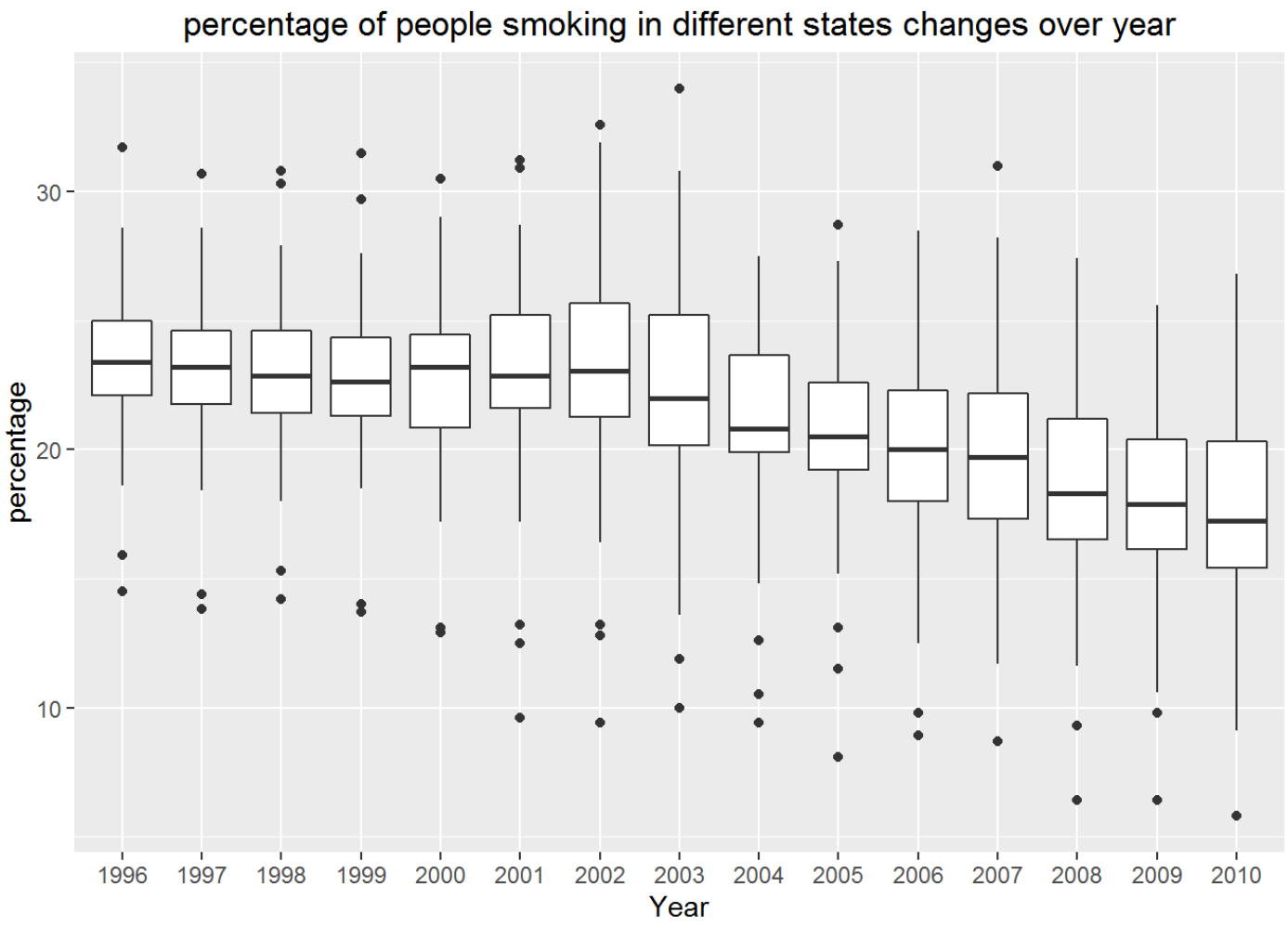
```
Tobacco.CU1f<-Tobacco.CU1f[ ,-c(7,9,10,11)]
# smoking status
Tobacco.CU1s<-Tobacco.CU1[Tobacco.CU1[,7]=="165SSA",]
Tobacco.CU1s<-Tobacco.CU1s[ ,-c(7,9,10,11)]
```

So far, We clean up the data. We later create a new dataset called “new.d” that each row contains one state in one year with all the percentage value that we have for each responce to each question based on the variables that we find out would influence the percentage value above. Now, we need to see the distribution of each dataset. First, we would show the distribution in current smoking dataset.

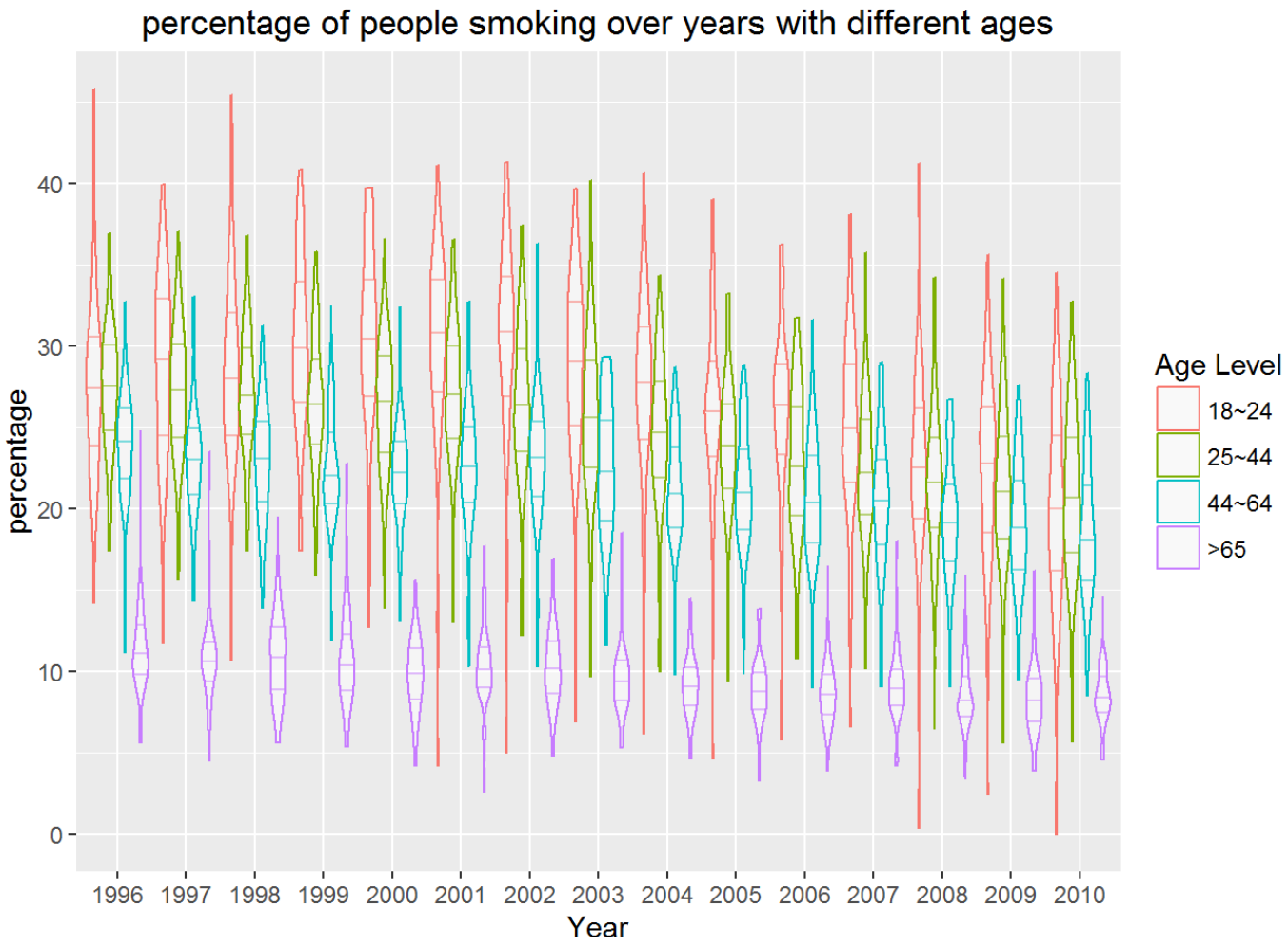
```
## Warning: package 'ggplot2' was built under R version 3.2.5
```



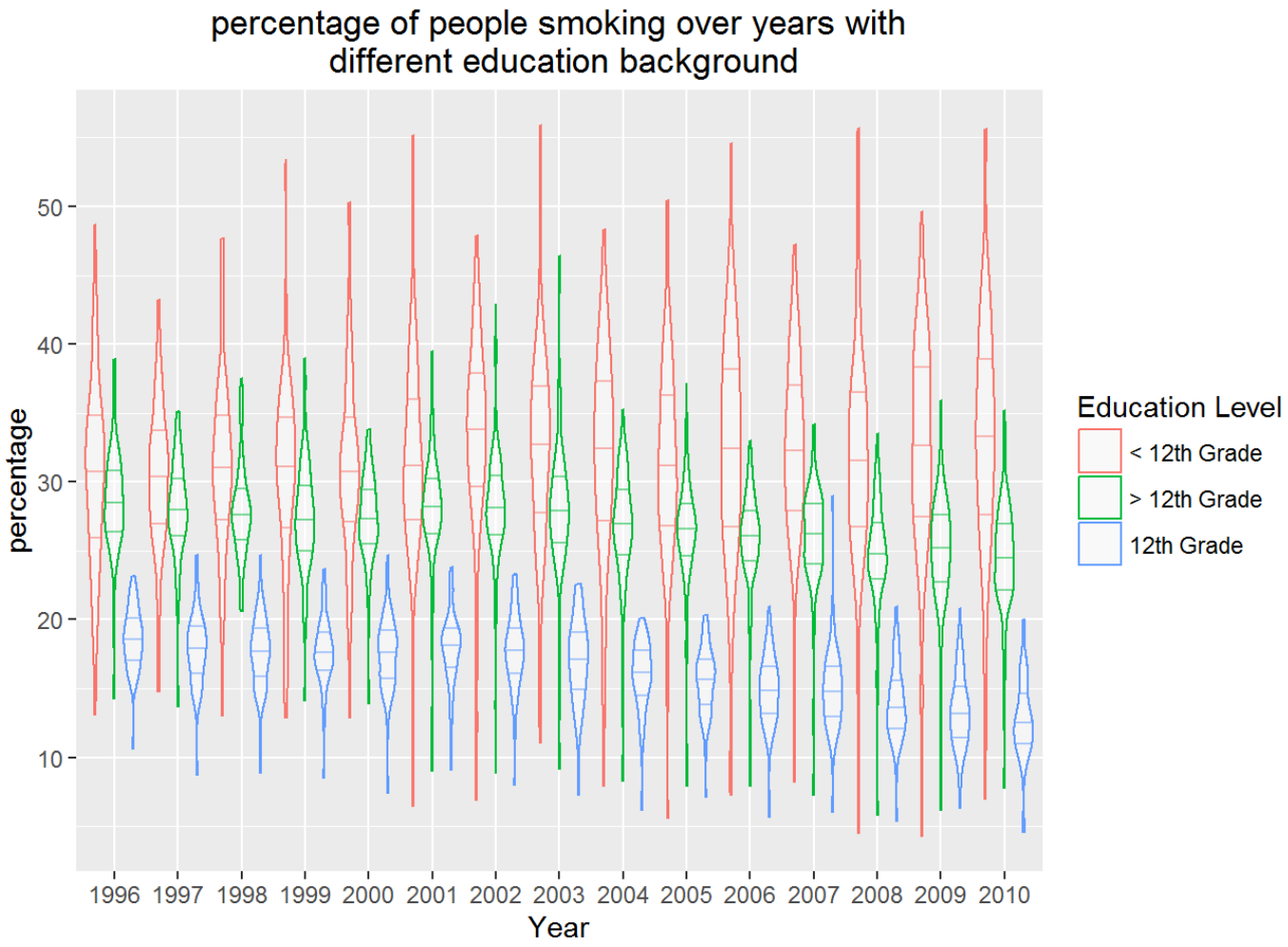
This plot is the distribution of the percentage of people currently smoke changes over state. We could tell from the plot that from 1996 to 2010, most states have 15 to 30 percentage people smoke. However, there are several states, such as “PR”, “UT” and “VI”, have much less percentage of people smoke. We could also see that for most states, percentage of people smoke is getting less over time. Now, we can figure out this trend with another plot.



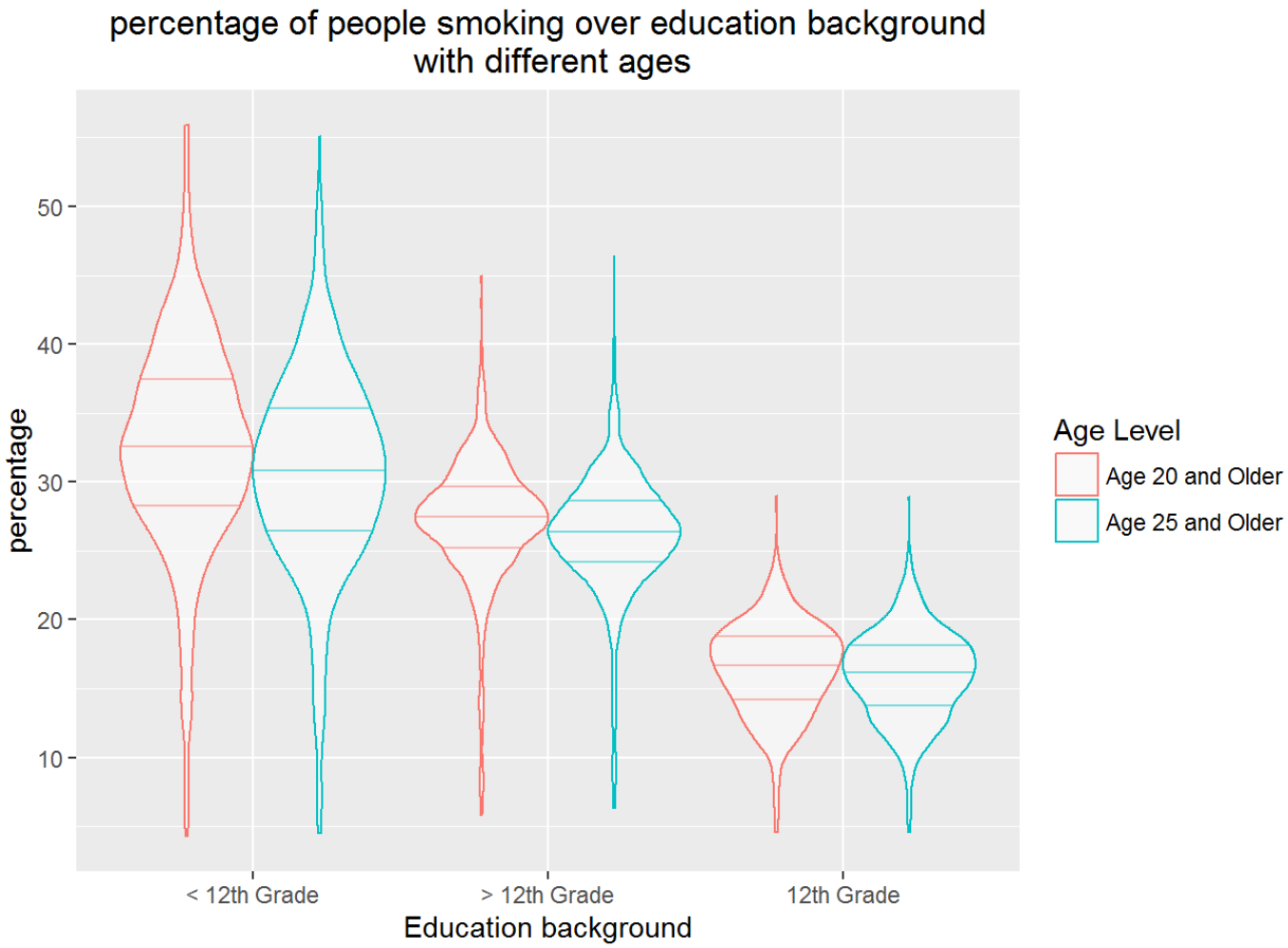
This plot is the distribution of the percentage of people currently smoke changes over time. It shows a trend that the percentage of people smoking over all states is getting less and less started from 2003.



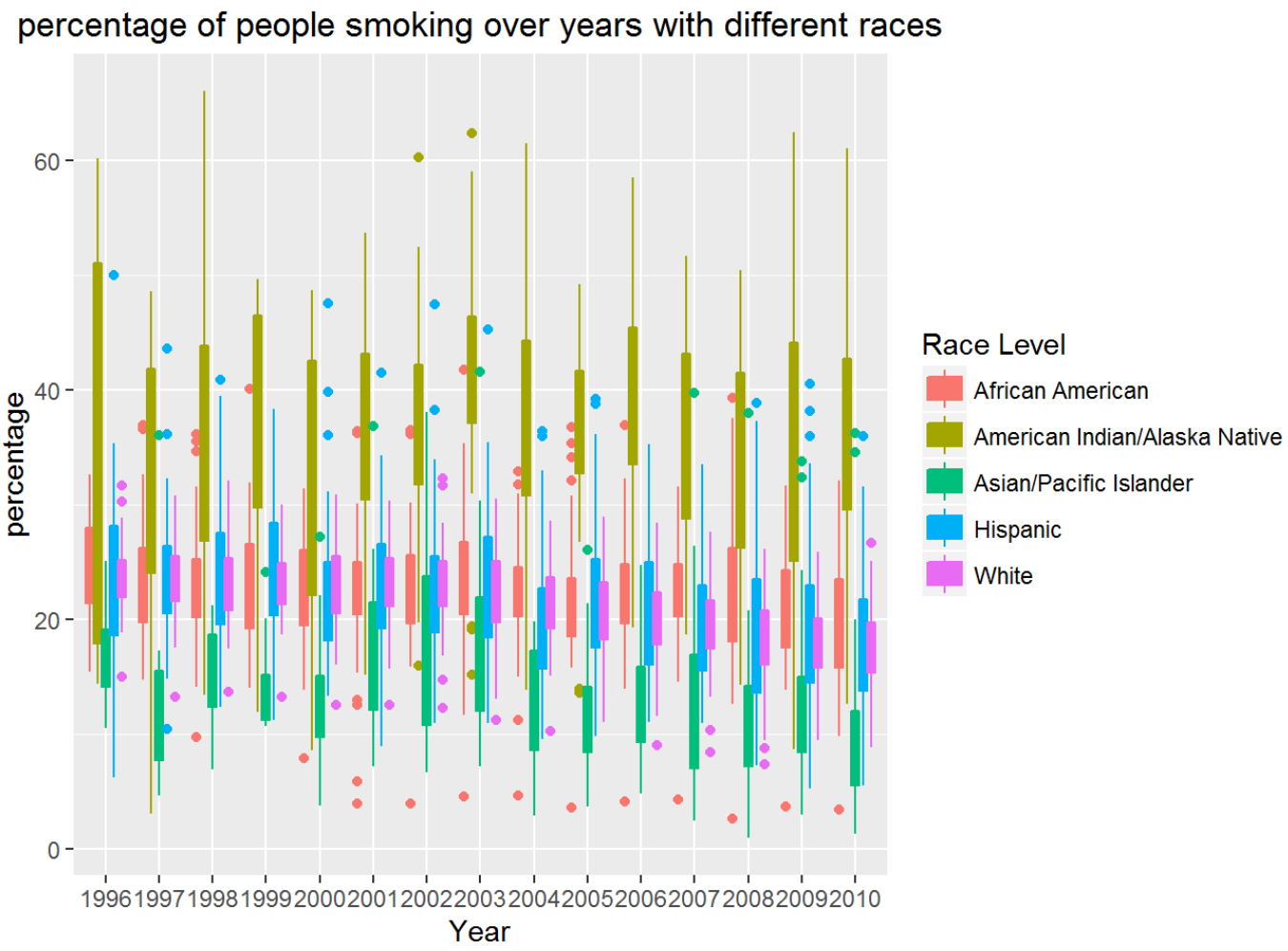
This plot is the distribution of the percentage of people currently smoke based on different age over year. We can see that age level “18~24” has a relatively higher percentage smoking and the age level over 65 has least percentage smoking.



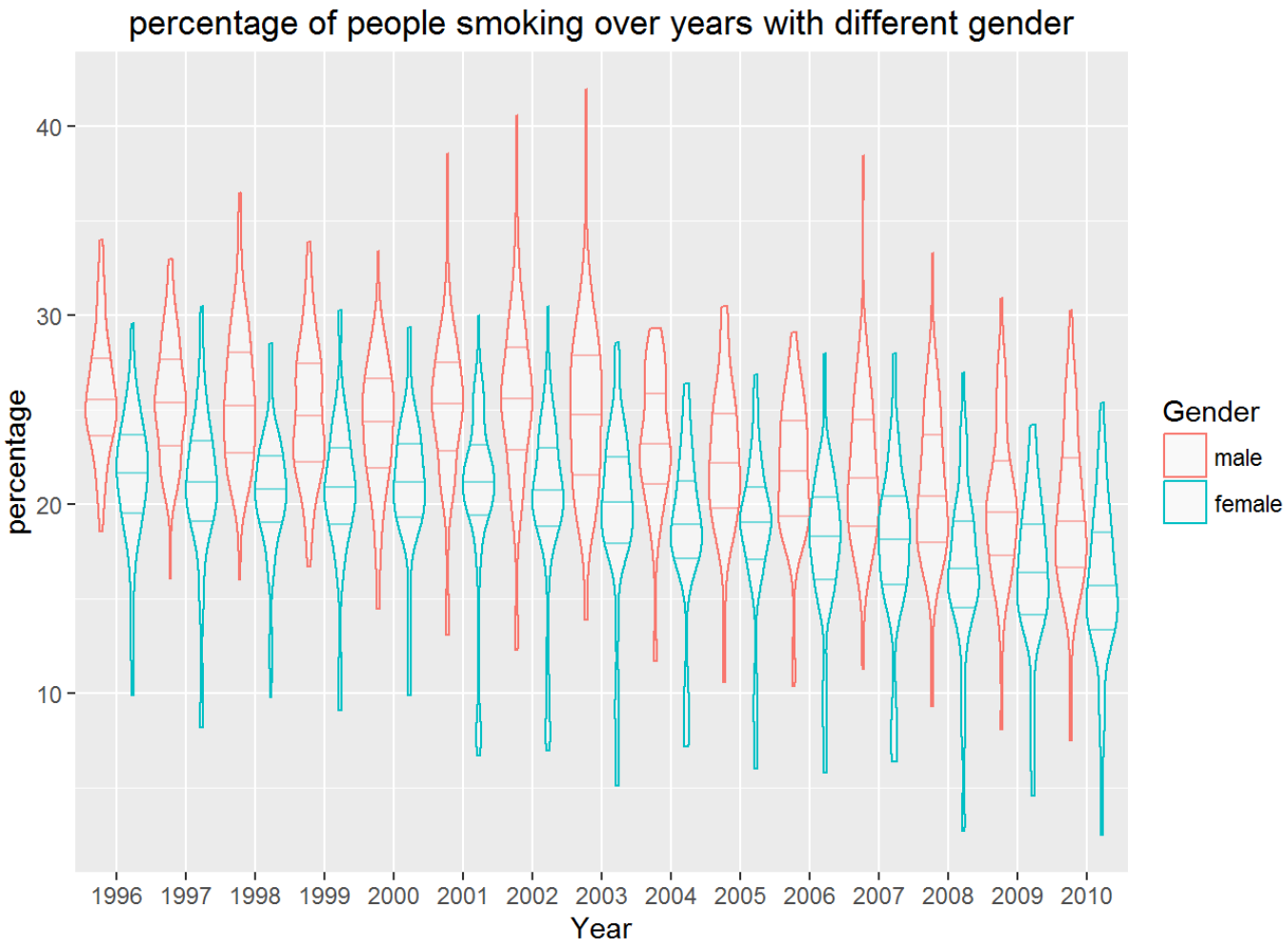
This plot is the distribution of the percentage of people currently smoke based on different education background over year. We can see that people with “<12th Grade” background would have the most percentage of people smoke. Whereas peole with “12th Grade” would have the least percentage of people smoke. This trend does not change much in 15 years.



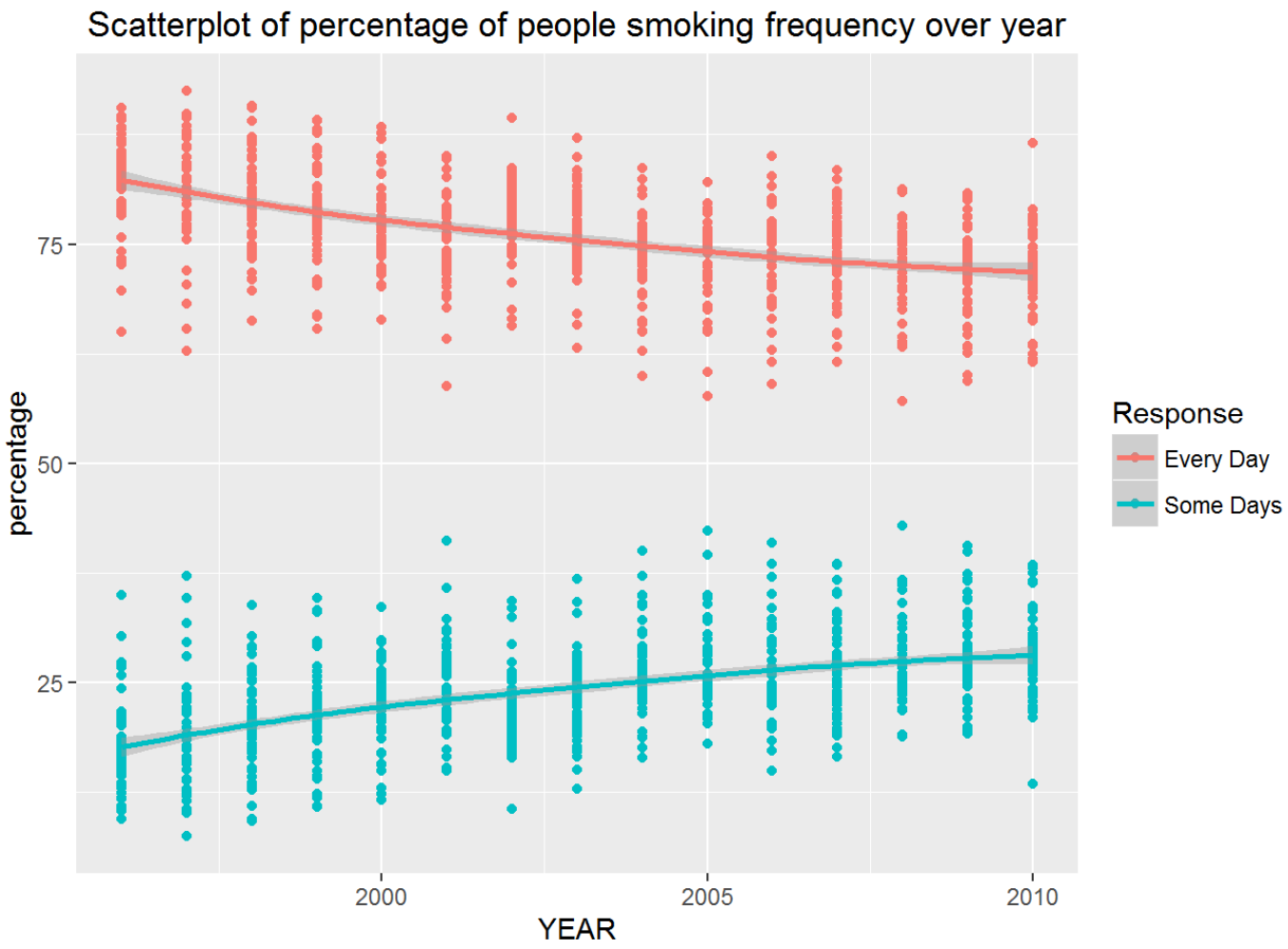
We also mentioned before that as the education background change, the age level would also change. Therefore, we make anthor plot to show how the age level would increase the percentage of people having the same education background. In this plot, we know that there is no much differences between two age level although the level “age 20 and older” would have a higher percentage of people smoke.



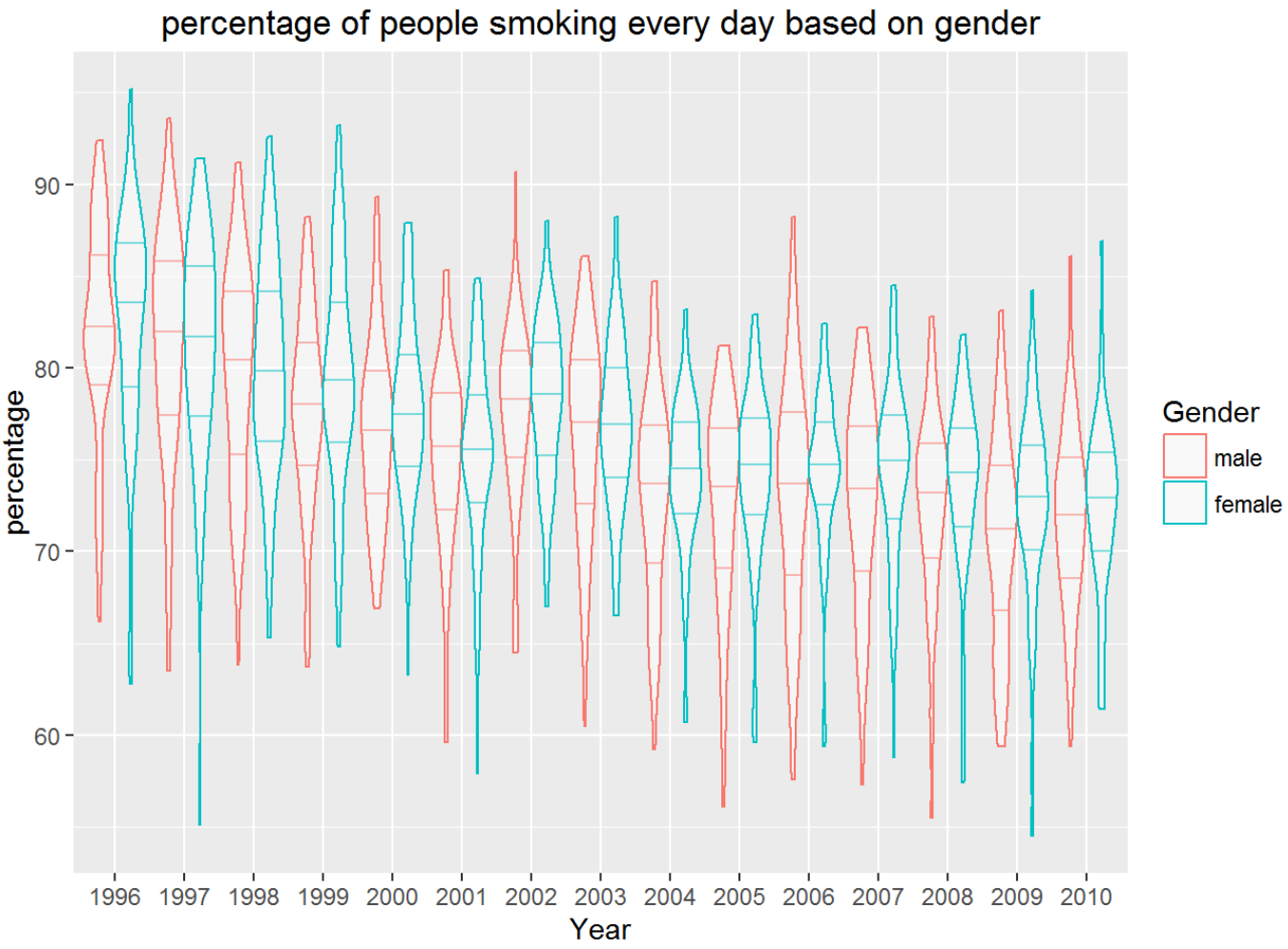
We could tell from this plot that there is no much change for the trend of people smoking based on different races over 15 years. We also could figure out that “Asian”" has least percentage of smoking and “American Indian/Alaska Native” has the most percentage of smoking.



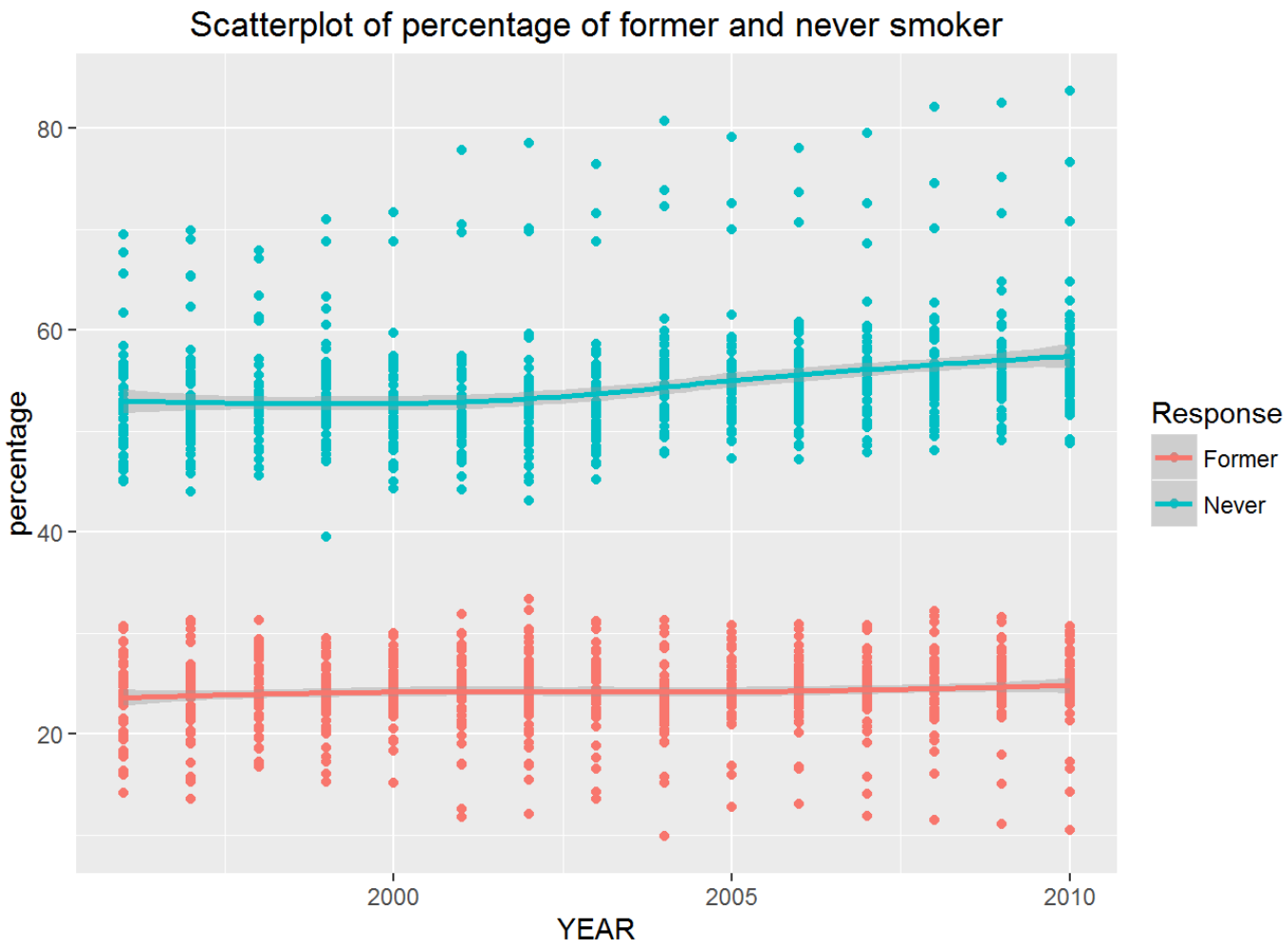
This plot shows the percentage of people smoking based on different gender over year. We could tell that we have less percentage of woman smoking than the percentage of man. The overall trend of percentage of people smoking also gets less over years.

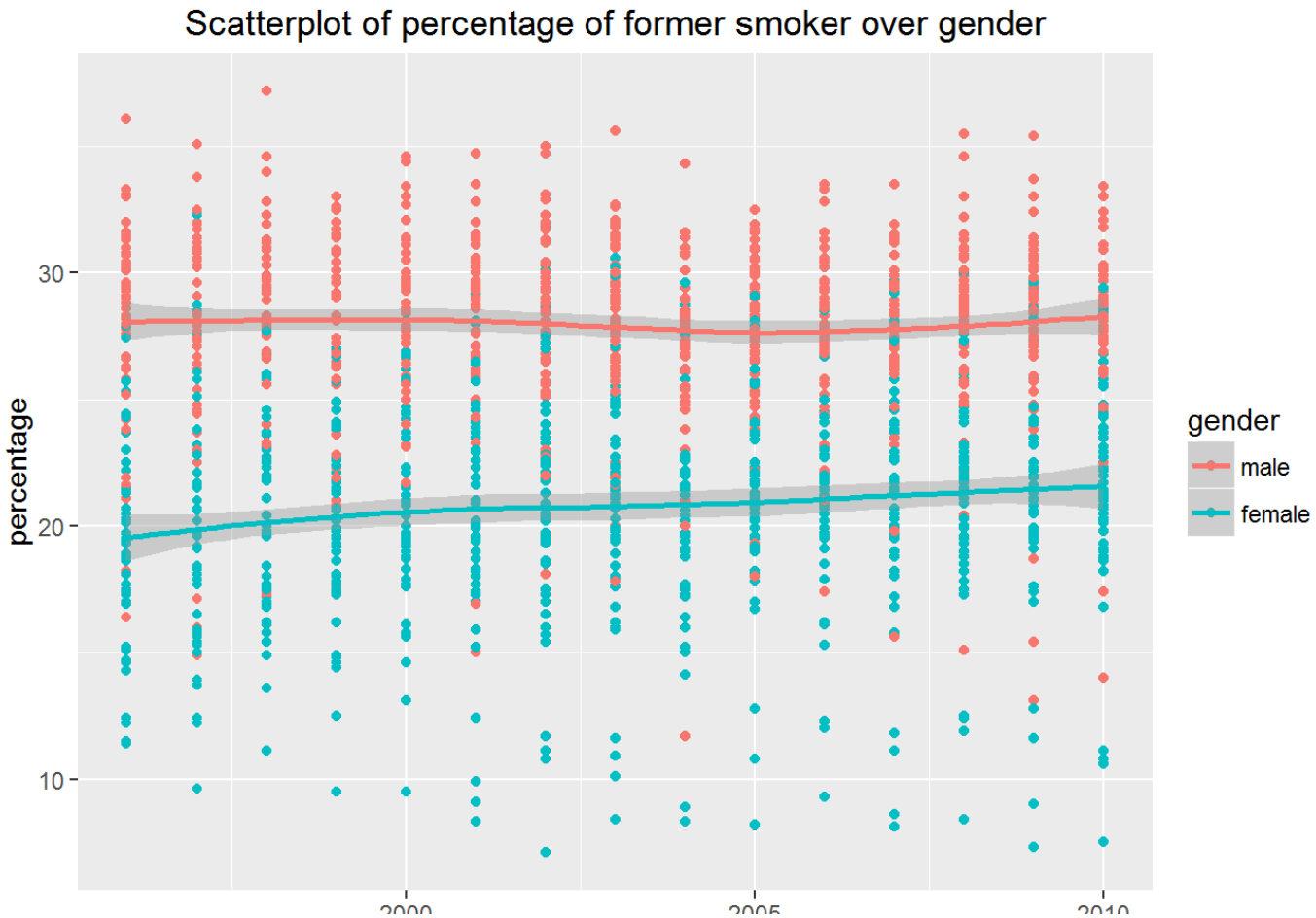


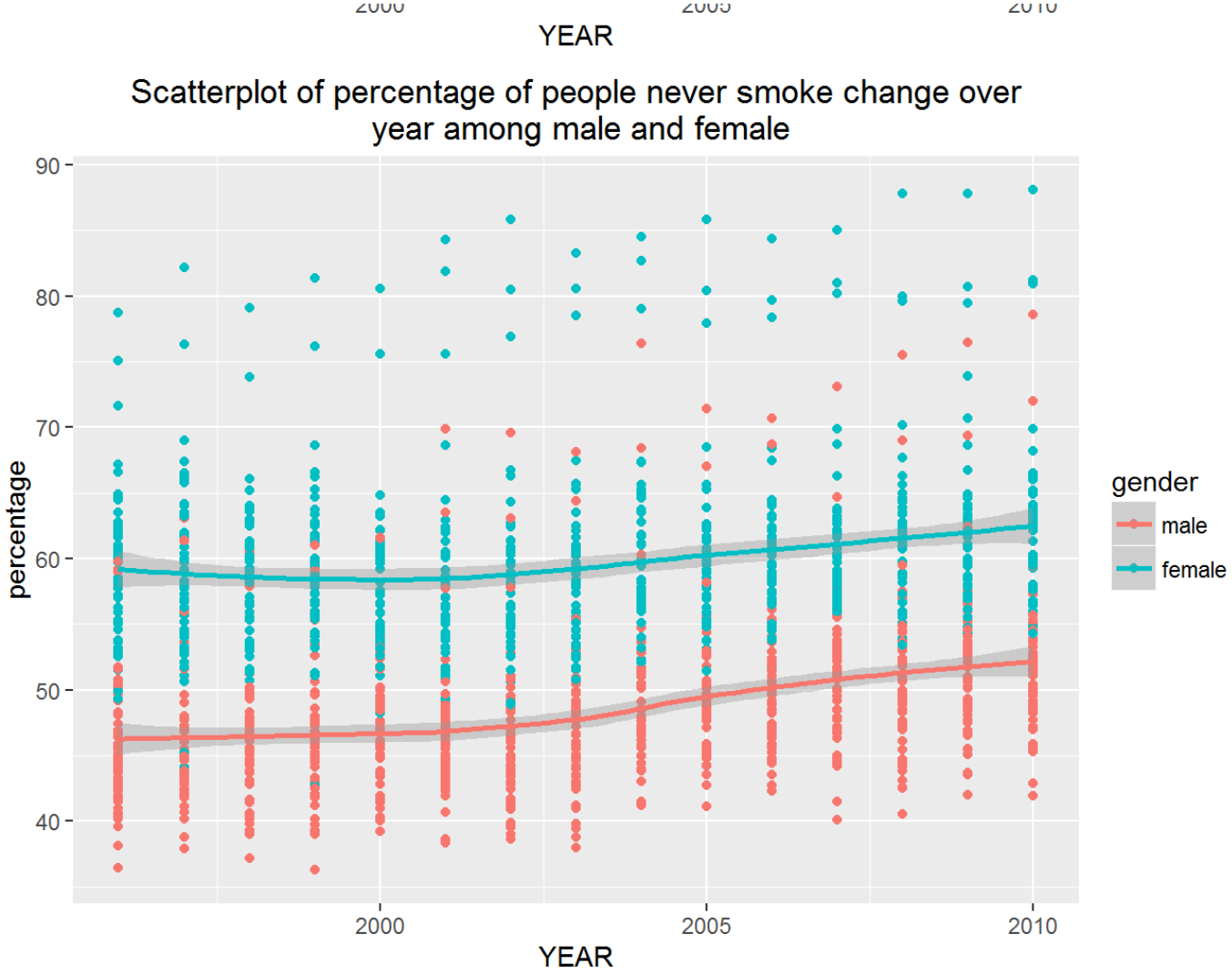
This plot shows the distribution of people smoke every day change over year from the smoking frequency dataset. We could tell that more and more percentage people smoke “Some days” instead of “Every Day”. However, the percentage of people choose to smoke every day is still higher than percentage of people choose to smoke some days.



This plot shows the distribution of people smoke every day change over year based on gender from the smoking frequency dataset. From the plot we could tell that the percentage of female smoking every day is slightly more than the percentage of male smoking every day.


Scatterplot of percentage of former and never smoker
percentage
80
60
40
20
2000
2005
2010
YEAR
Response
Former
Never


Scatterplot of percentage of former smoker over gender
percentage
30
20
10
gender
male
female



The above three plots show the distribution of two different reponses from smoking status dataset. We choose not to do the “current” response since we have already done that in the previous part about current smoking dataset. We could tell from the plots that the percentage of male that are former smokers is higher than the one of female. Also more percentage of female never smoke. The overall trend does not change much over years.

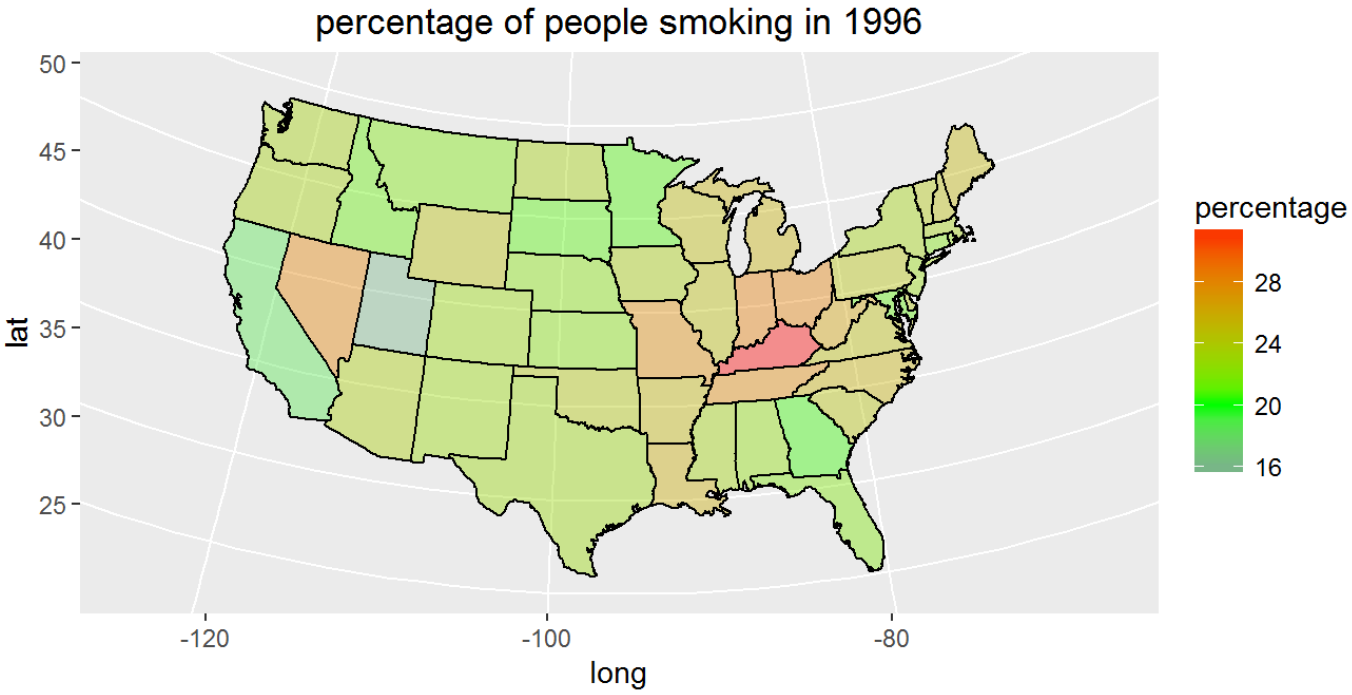
Results

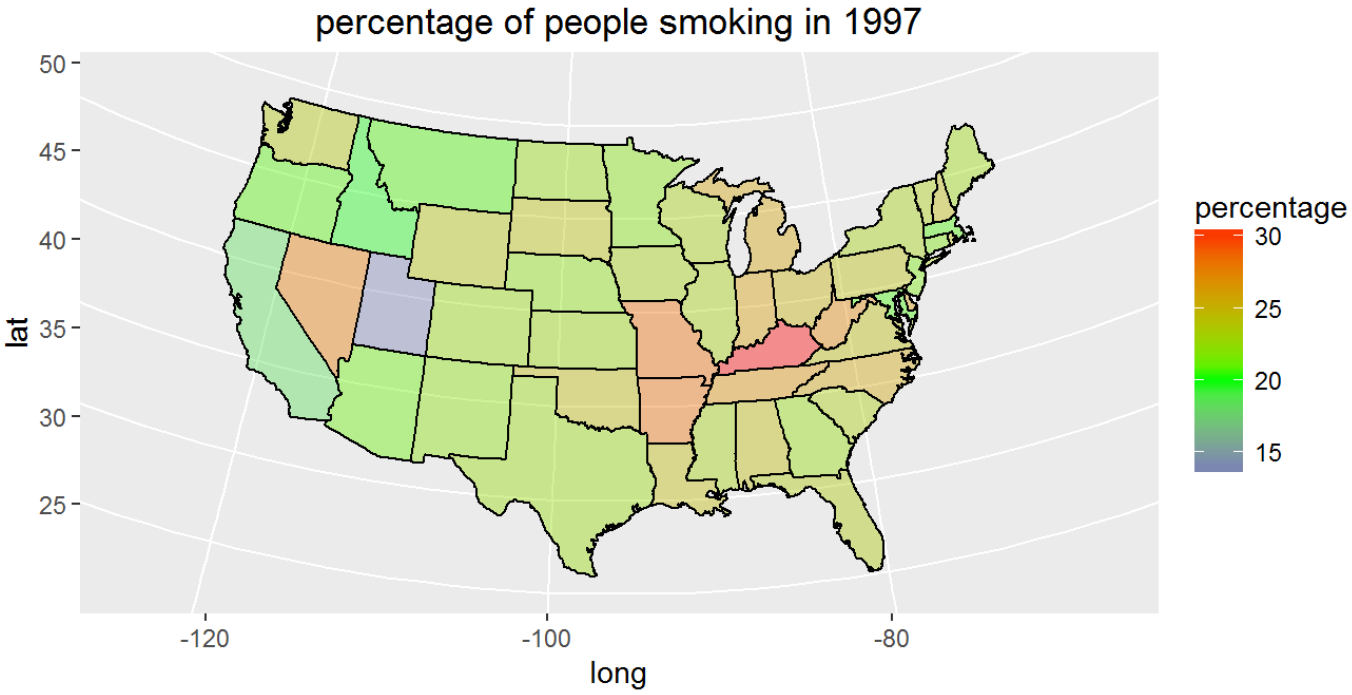
In this part, we are gonna show our observations and what we found from these observations. We have created 5 groups of visualization using ggmap. Each group of observation shows a 15 year change of smoking trend based on the new dataset called “new.d” in which each row represent one state in one year with all the percentage of each response for each question.

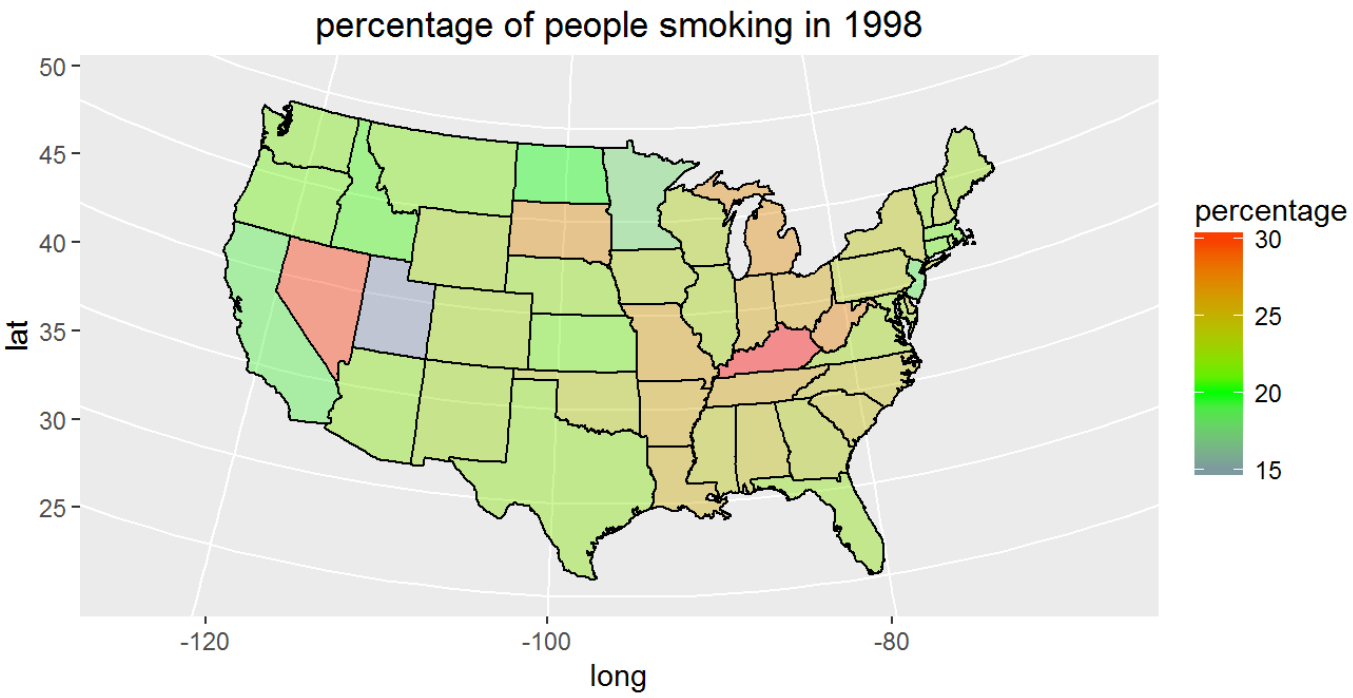
The first group of observations present current smoking for overall gender, race, education level and age from 1996 to 2010.

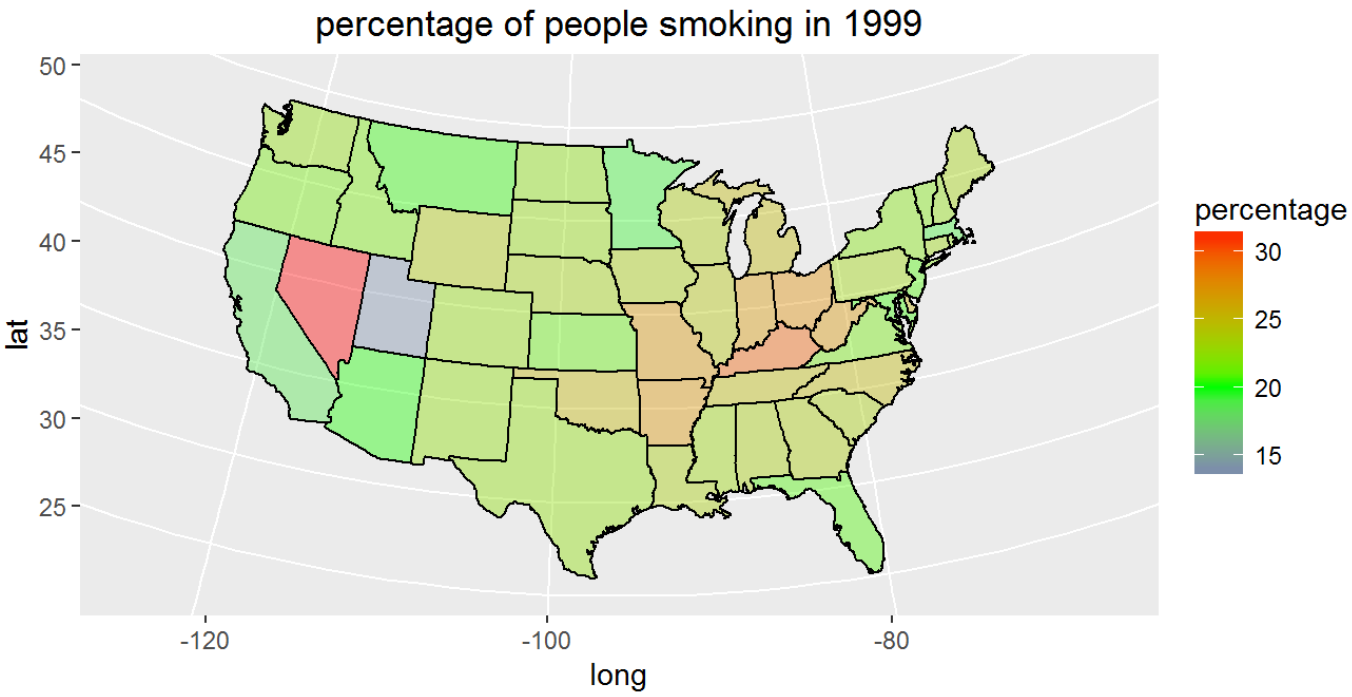
```
## Warning: package 'maps' was built under R version 3.2.5
```

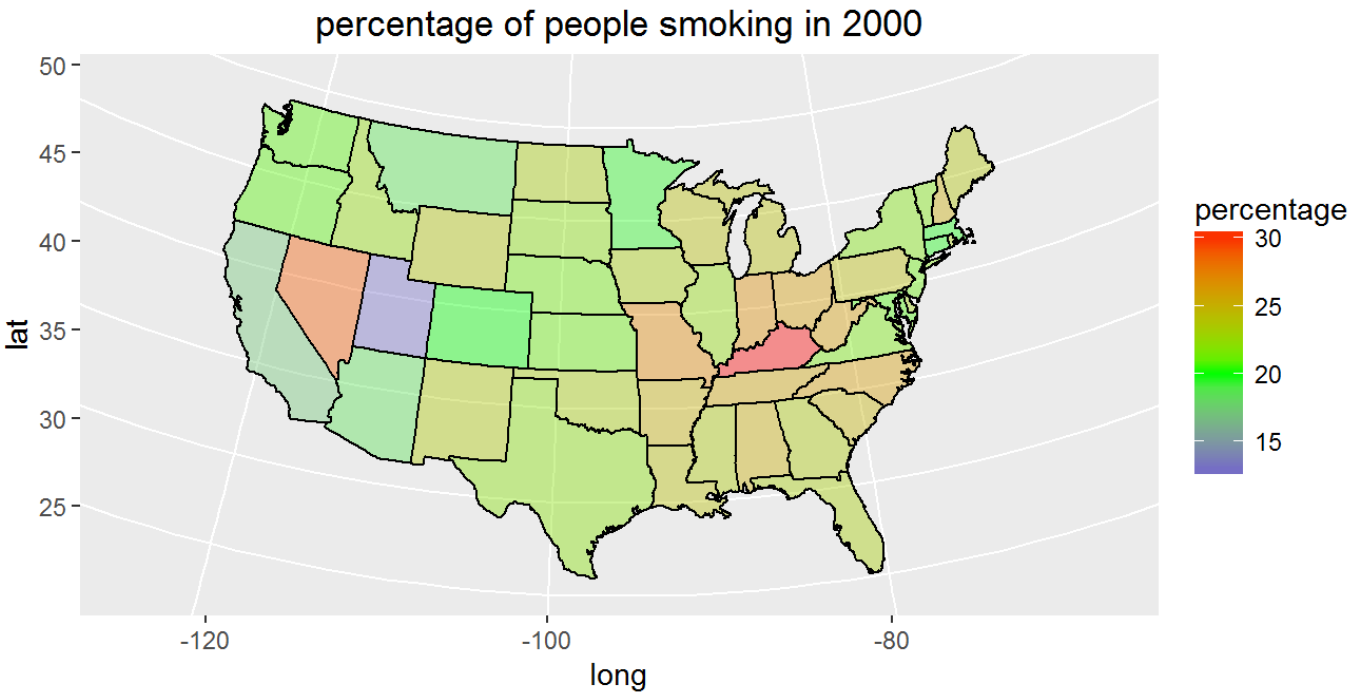
```
##
##  # maps v3.1: updated 'world': all lakes moved to separate new #
##  # 'lakes' database. Type '?world' or 'news(package="maps")'.  #
```

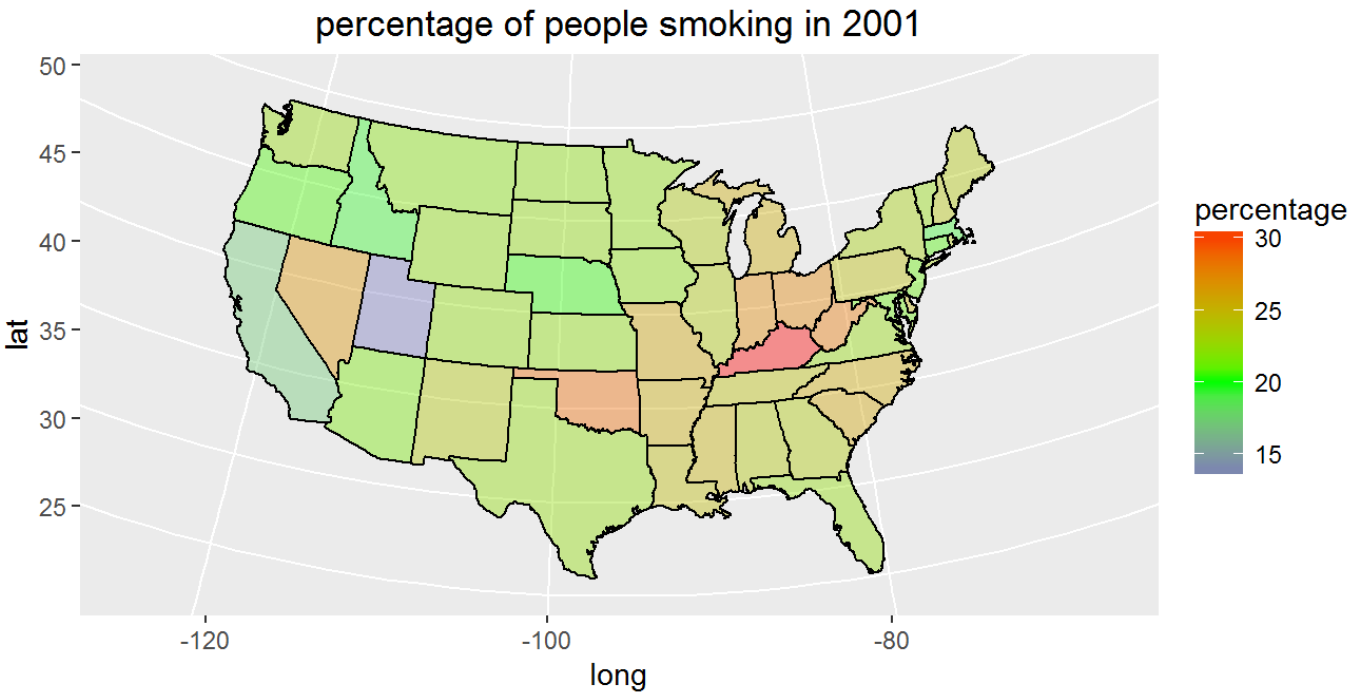

percentage of people smoking in 1996
lat
long
50
45
40
35
30
25
-120
-100
-80
percentage
28
24
20
16

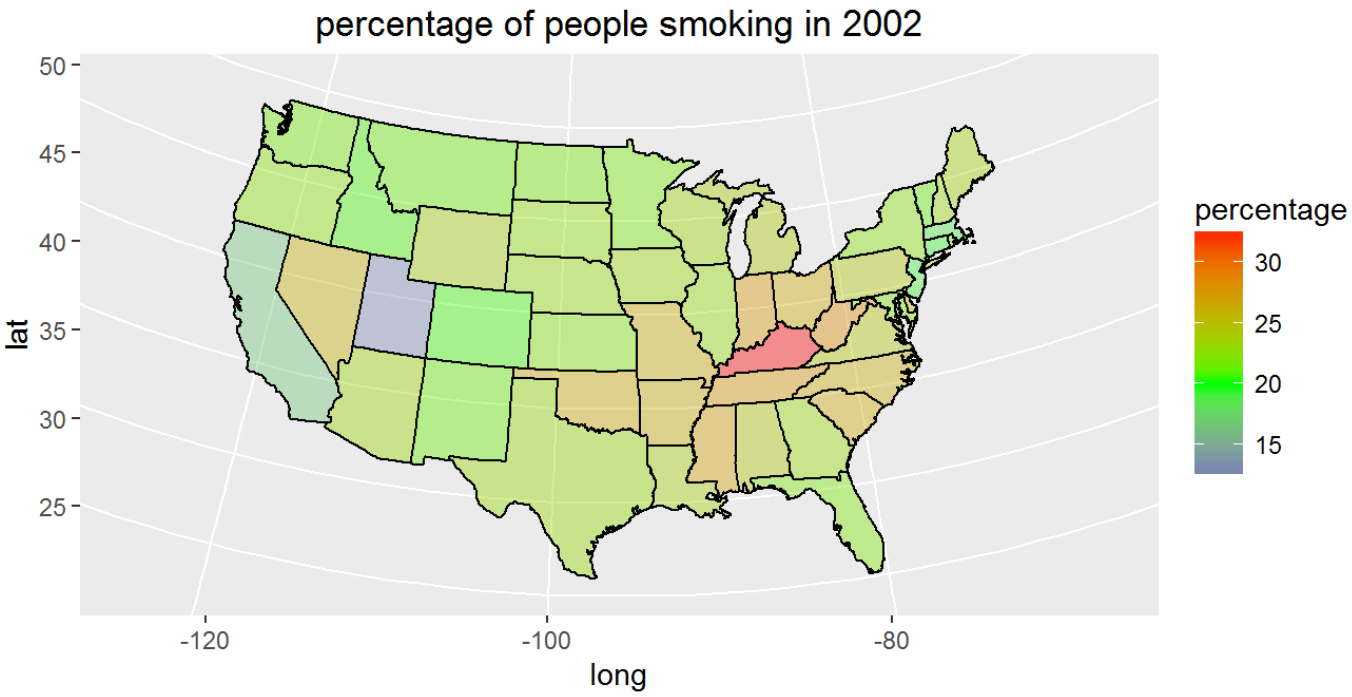

percentage of people smoking in 1997
lat
long
50
45
40
35
30
25
-120
-100
-80
percentage
30
25
20
15

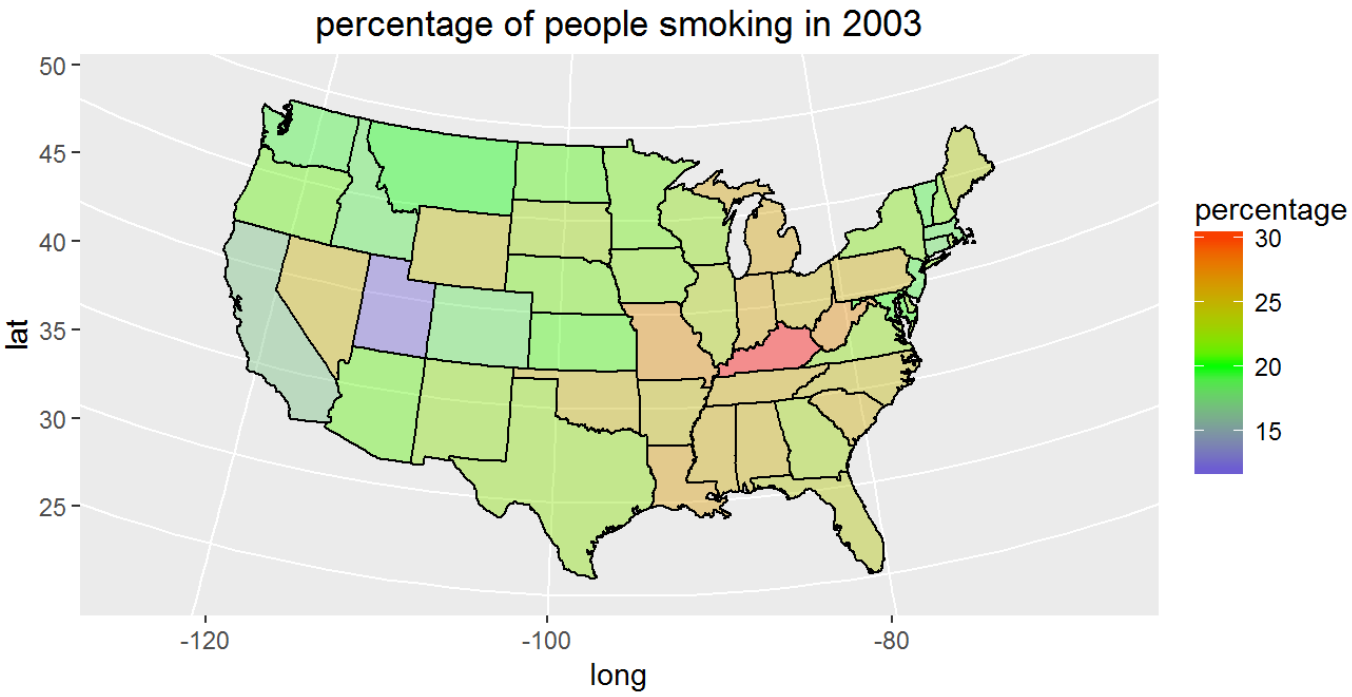

percentage of people smoking in 1998
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

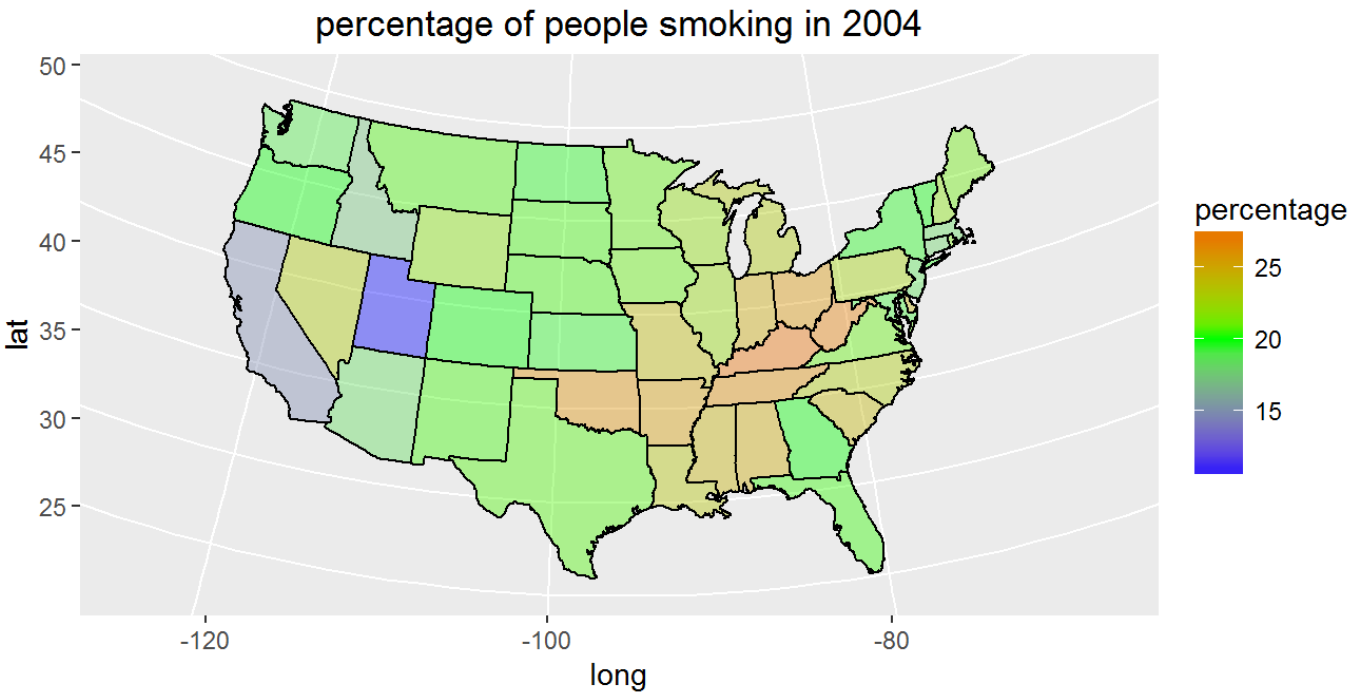

percentage of people smoking in 1999
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

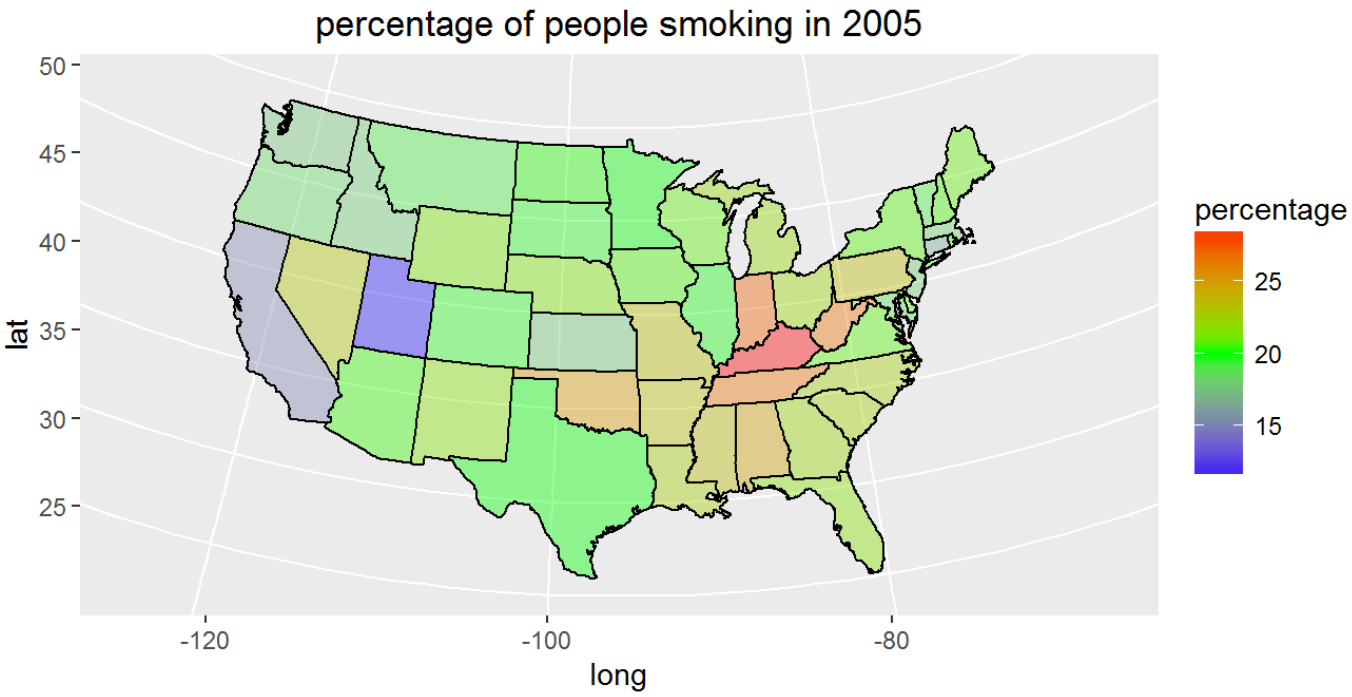

percentage of people smoking in 2000
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

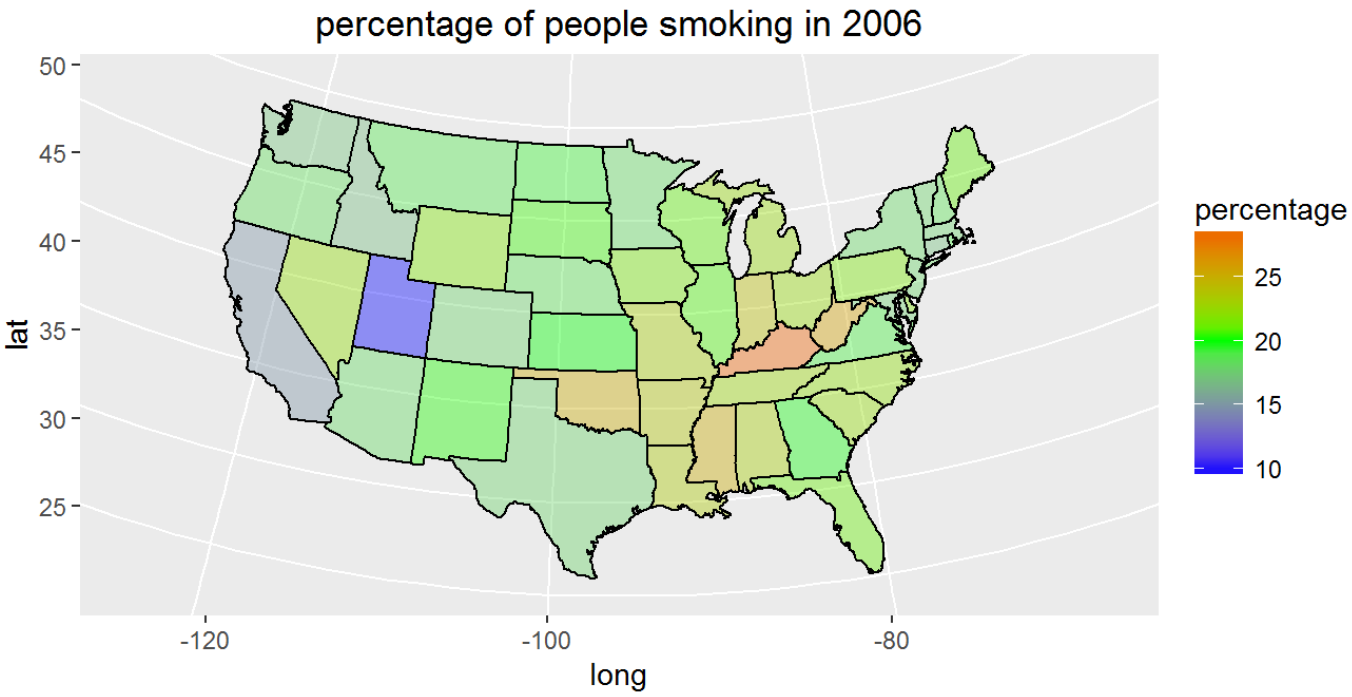

percentage of people smoking in 2001
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

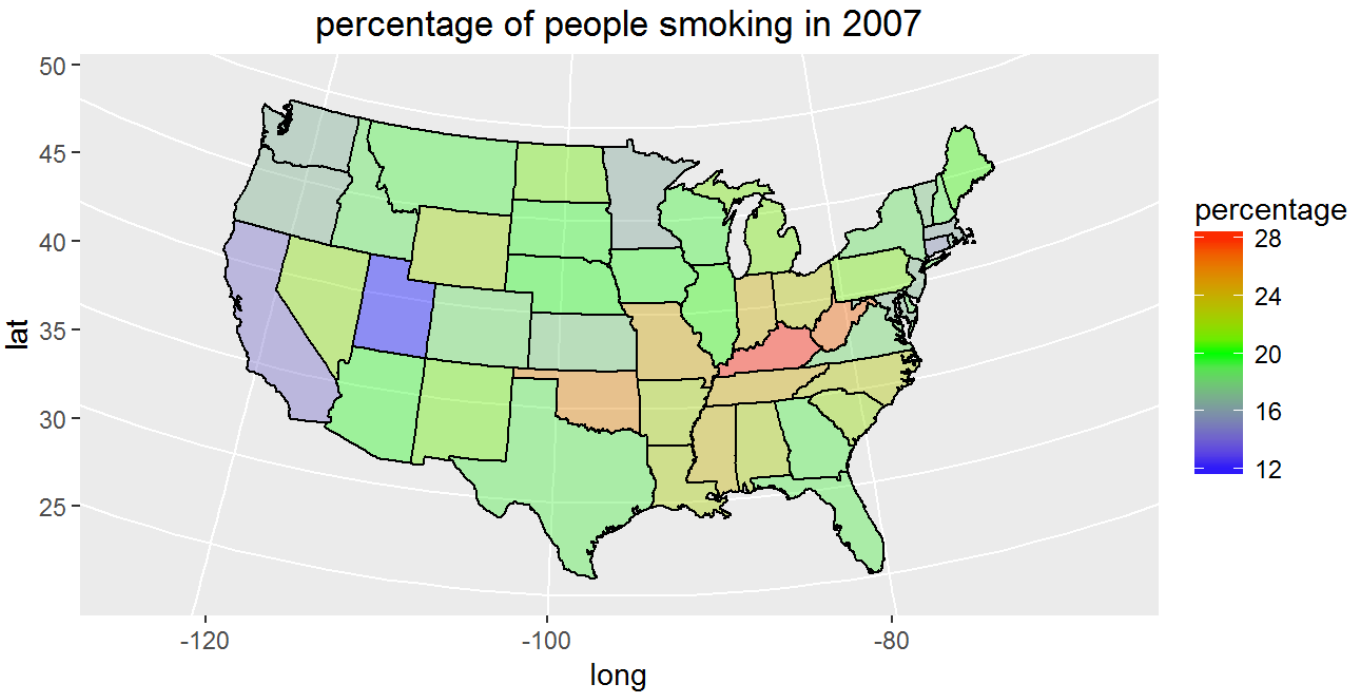

percentage of people smoking in 2002
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

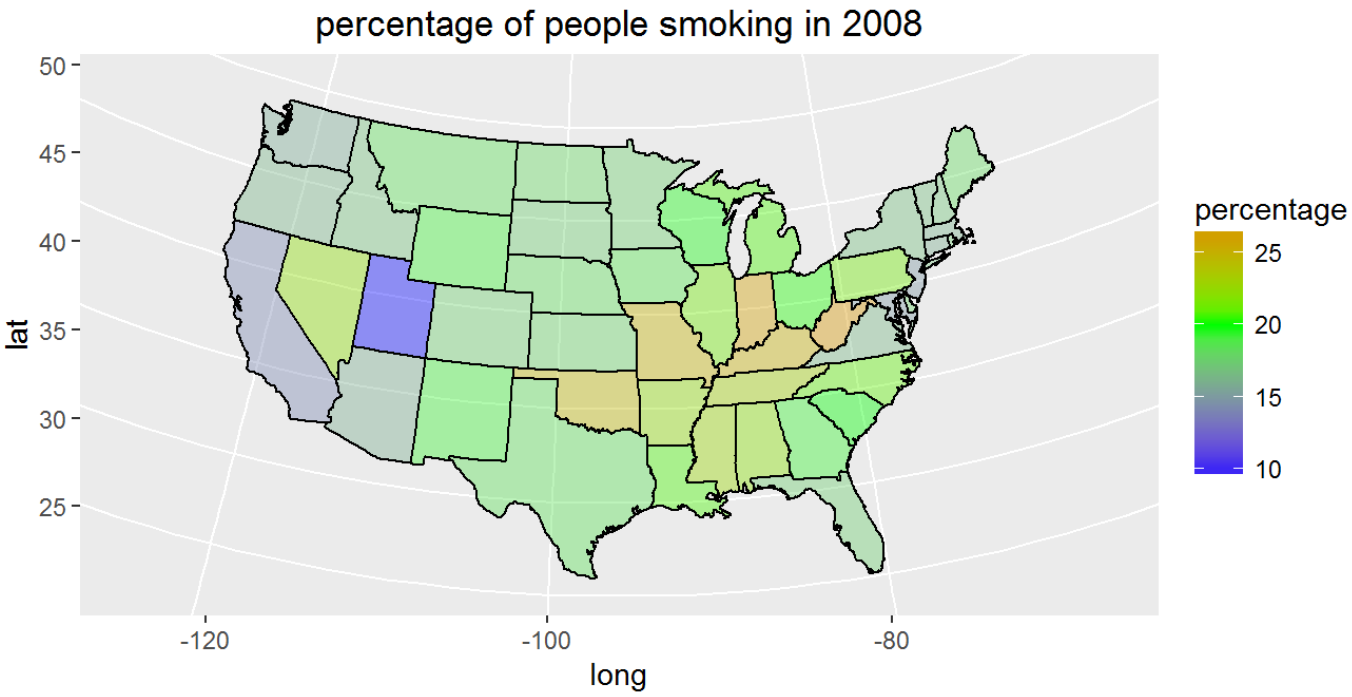

percentage of people smoking in 2003
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
30
25
20
15

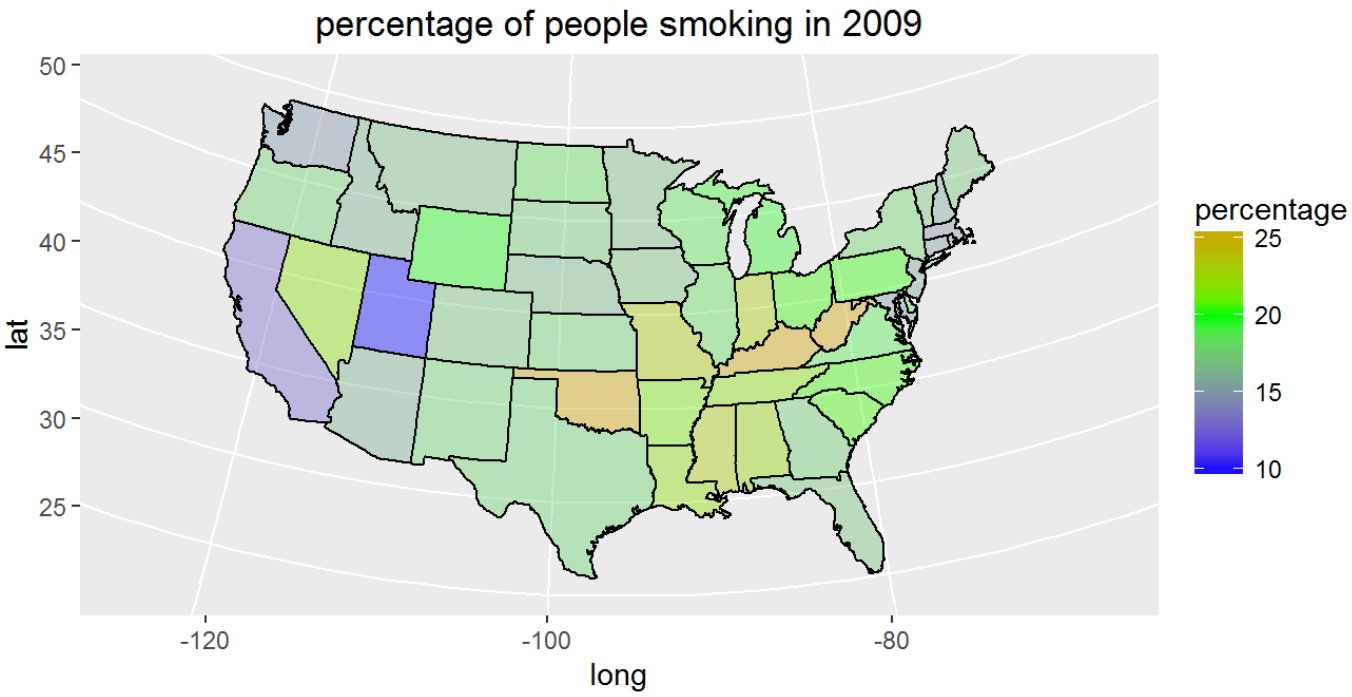

percentage of people smoking in 2004
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
25
20
15

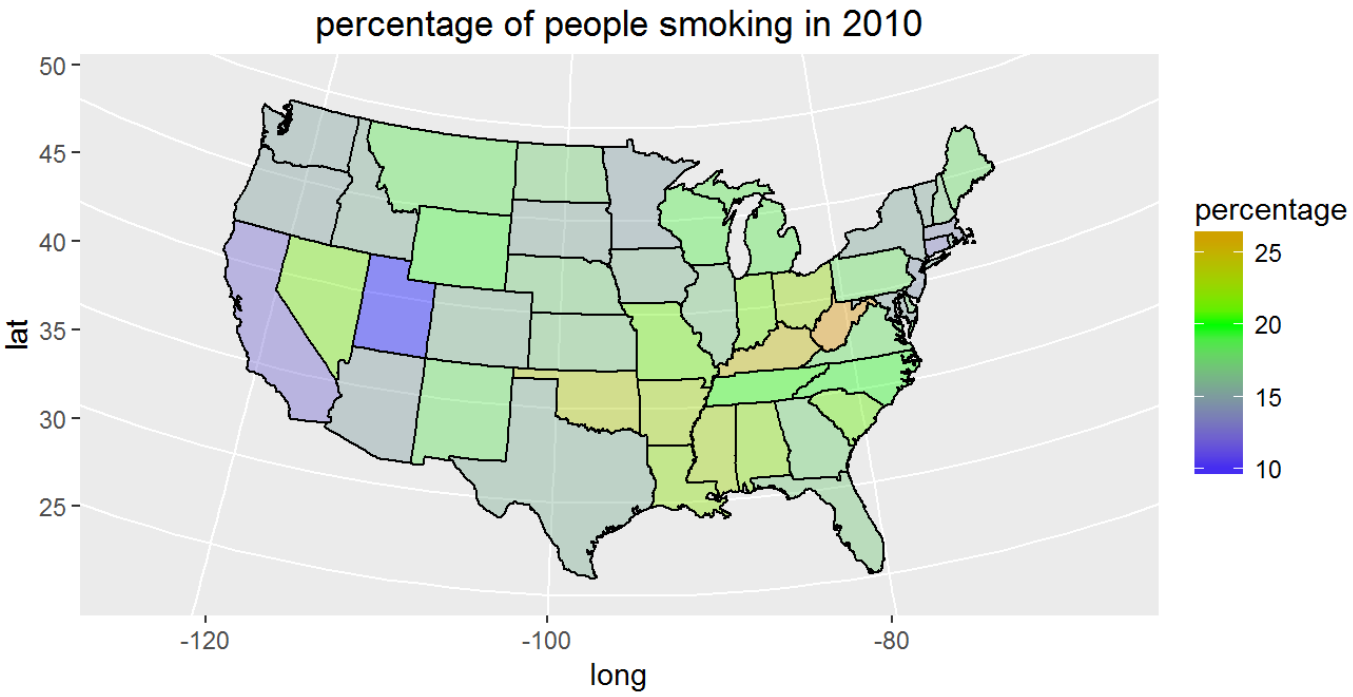

percentage of people smoking in 2005
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
25
20
15


percentage of people smoking in 2006
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
25
20
15
10


percentage of people smoking in 2007
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
28
24
20
16
12


percentage of people smoking in 2008
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
25
20
15
10

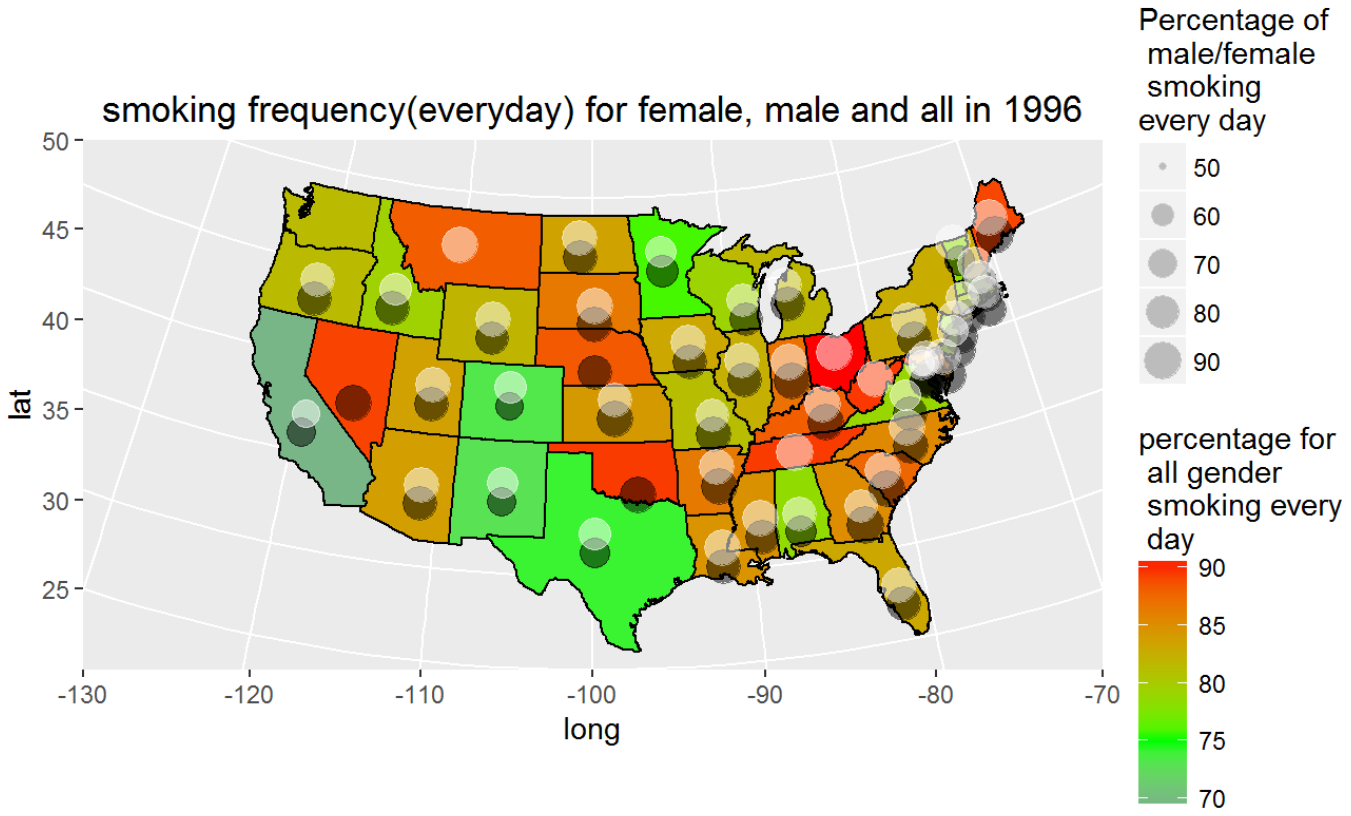

percentage of people smoking in 2009
50
45
40
35
30
25
lat
-120
-100
-80
long
percentage
25
20
15
10



The overall trend for the percenge of people smoking decreases over 15 years. According to the plots we have, we could tell that State “NV” has a relatively high percentage of people smoking in 1990s and later the percentage of people smoking gets down in 2001. However, it is still higher than the lots of states except state “KY”. Whereas, state “KY” always keeps a high percentage of people smoking over this 15-year-period. It has a lower percentage in 2008 and being exceeded by state “NV” in 1999. The interesting thing is that state “UT” has a lowest percentage of people smoking among all the states in 15-year-period.

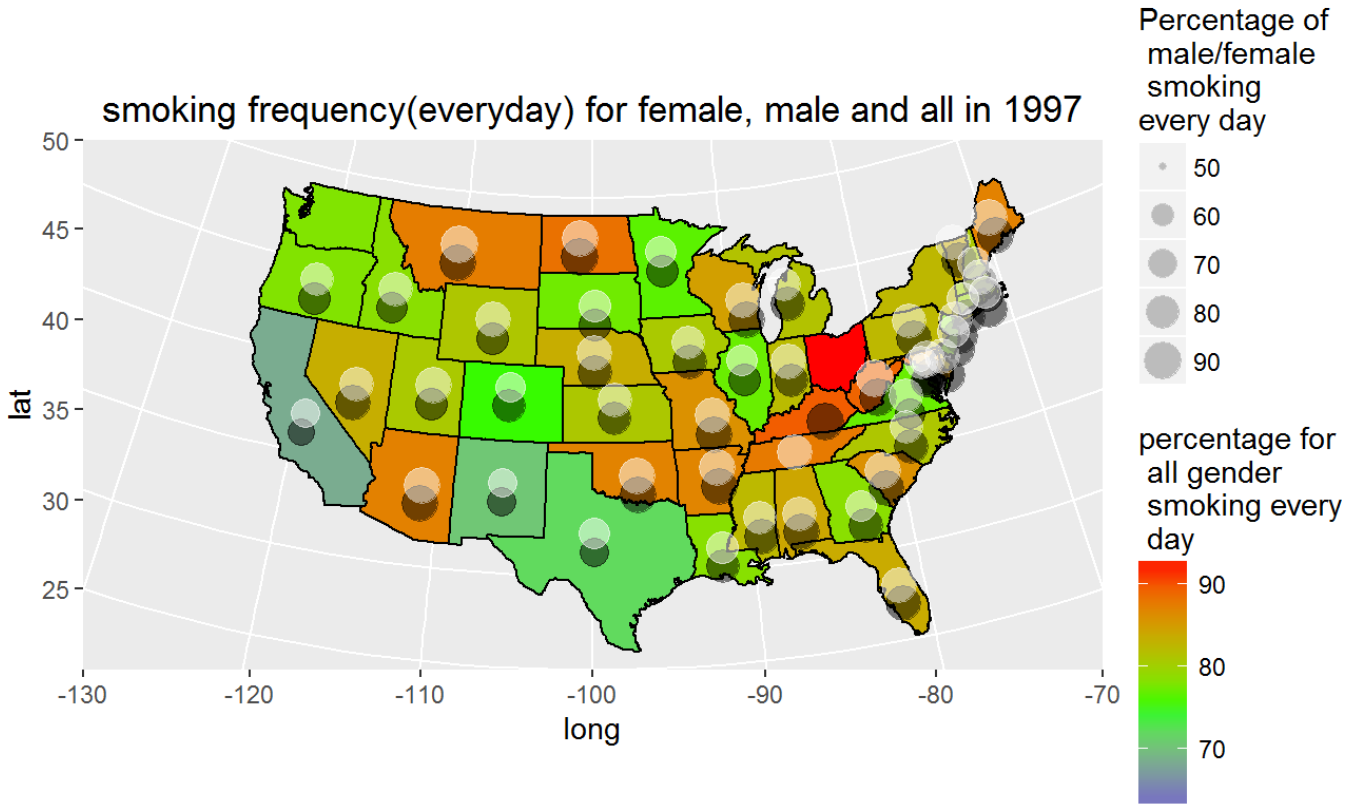
```
## Warning: Removed 7 rows containing missing values (geom_point).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```

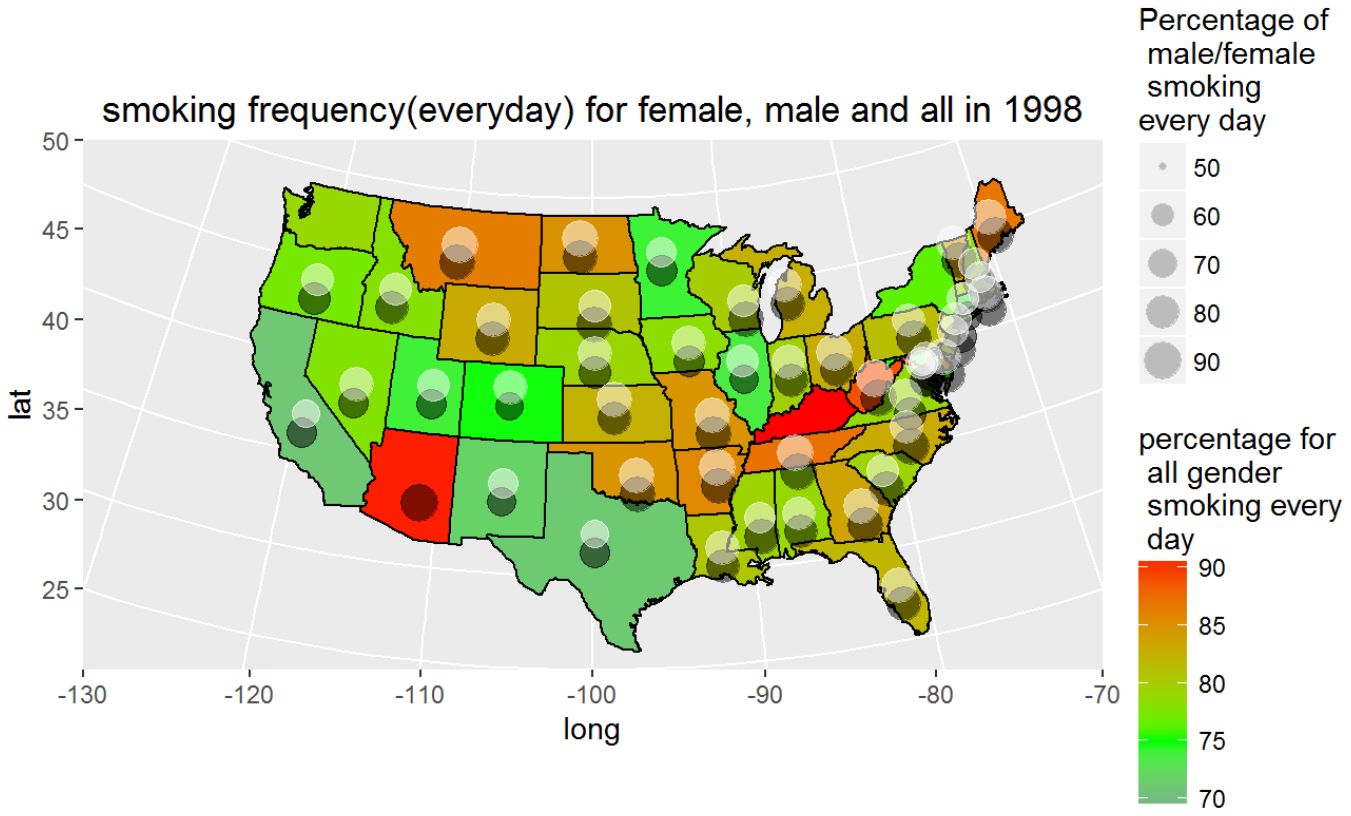



```
## Warning: Removed 5 rows containing missing values (geom_point).
```

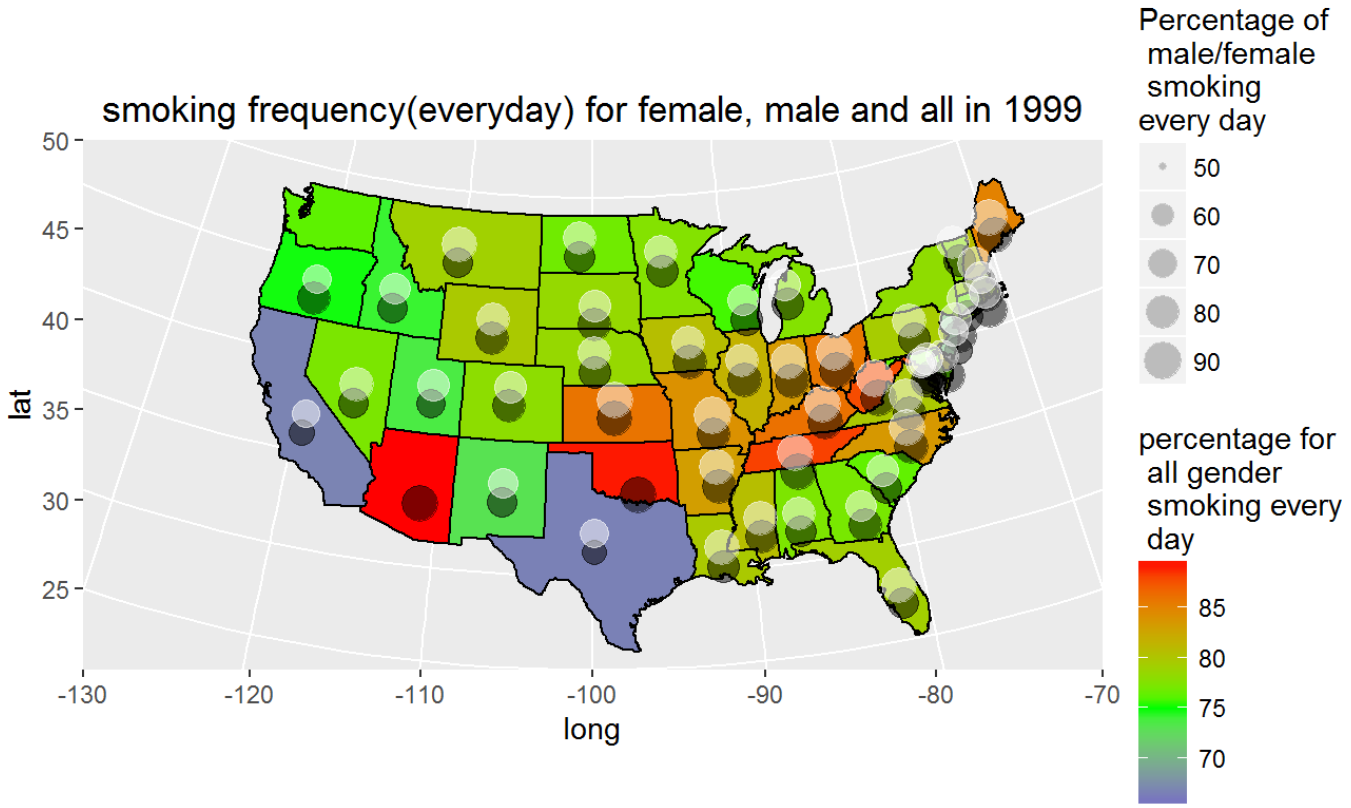
```
## Warning: Removed 5 rows containing missing values (geom_point).
```




```
## Warning: Removed 4 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```

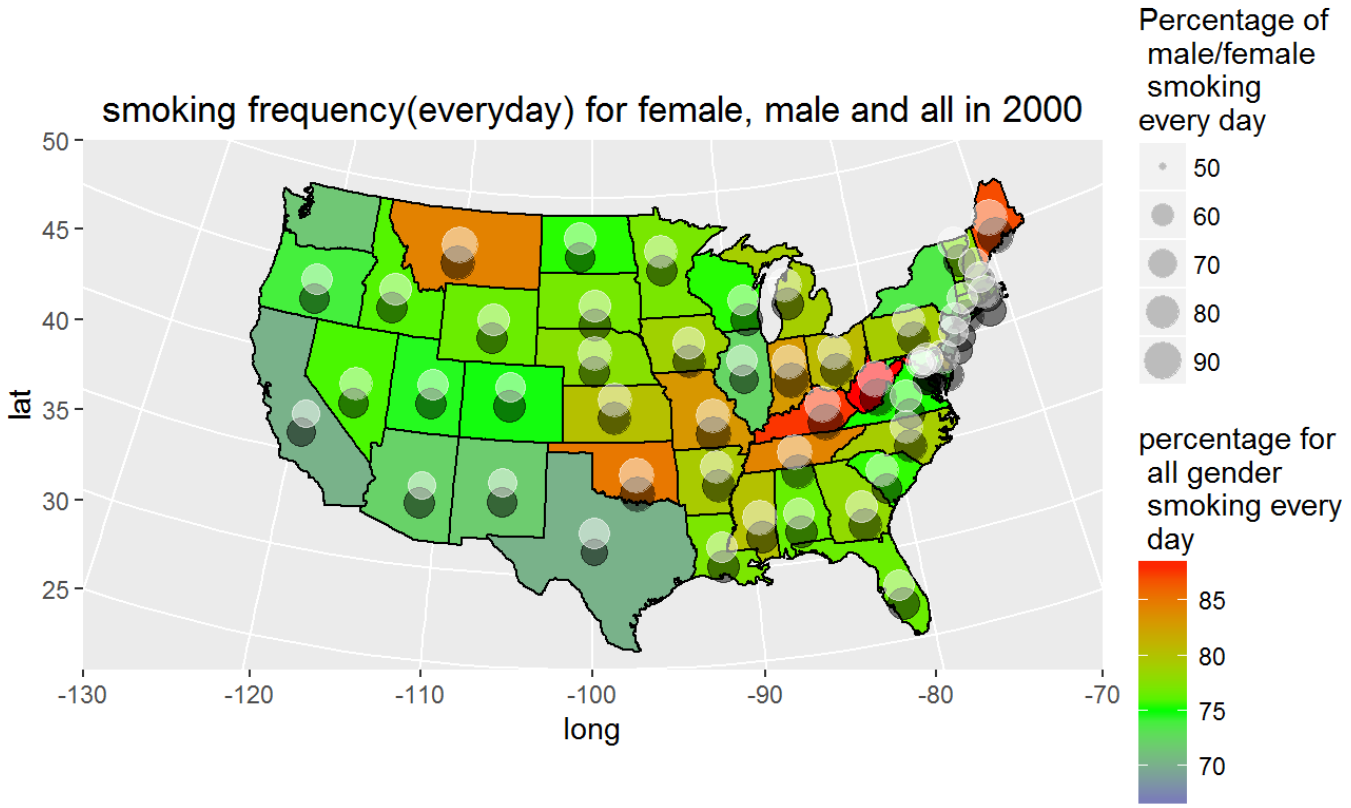



```
## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```



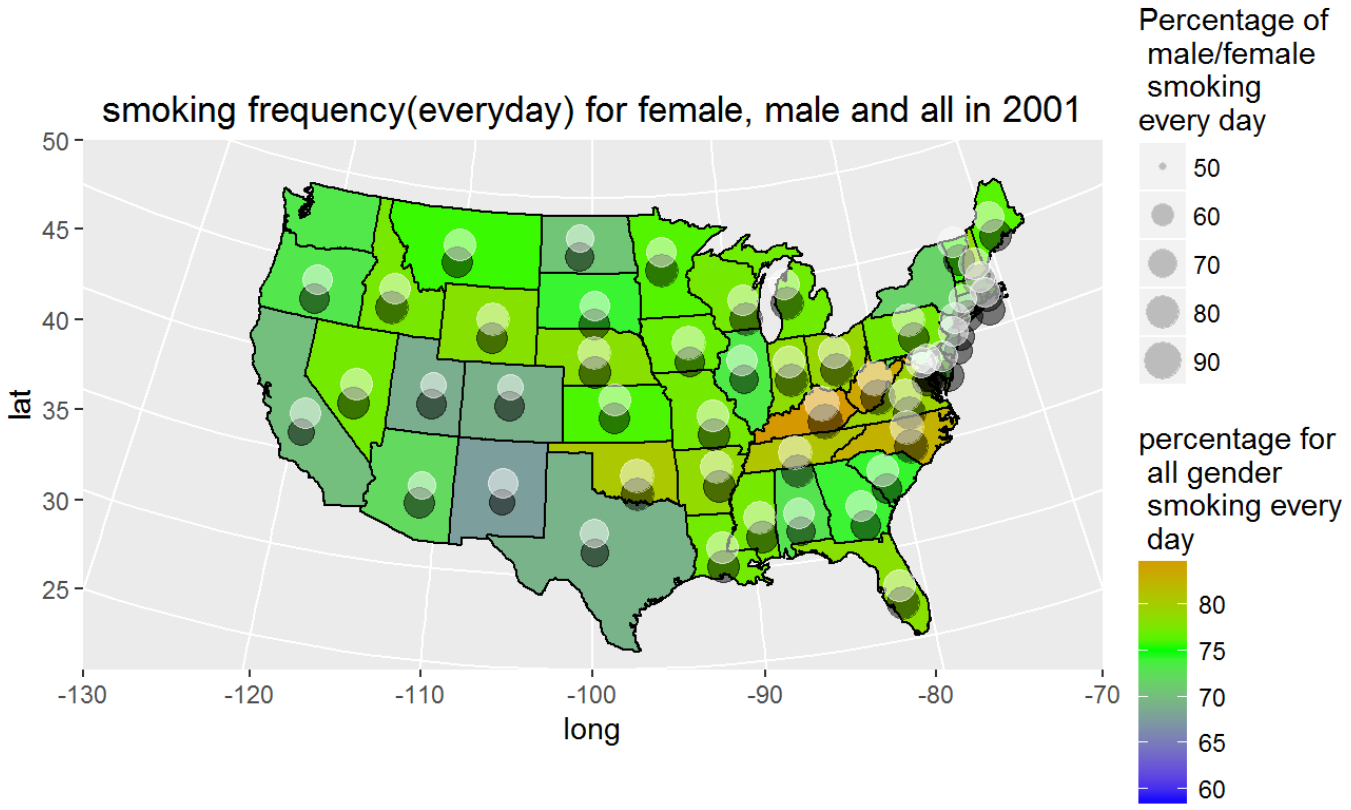

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

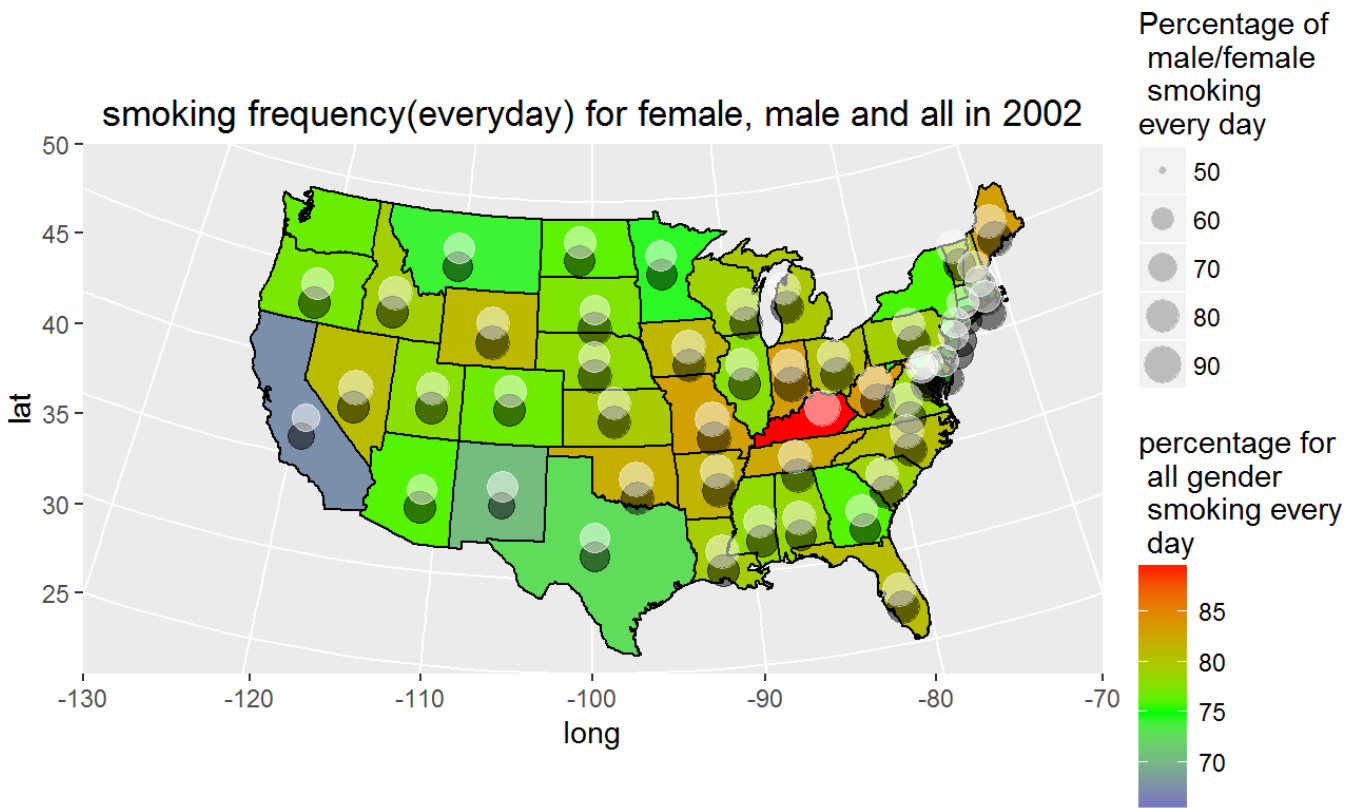



```
## Warning: Removed 1 rows containing missing values (geom_point).
```

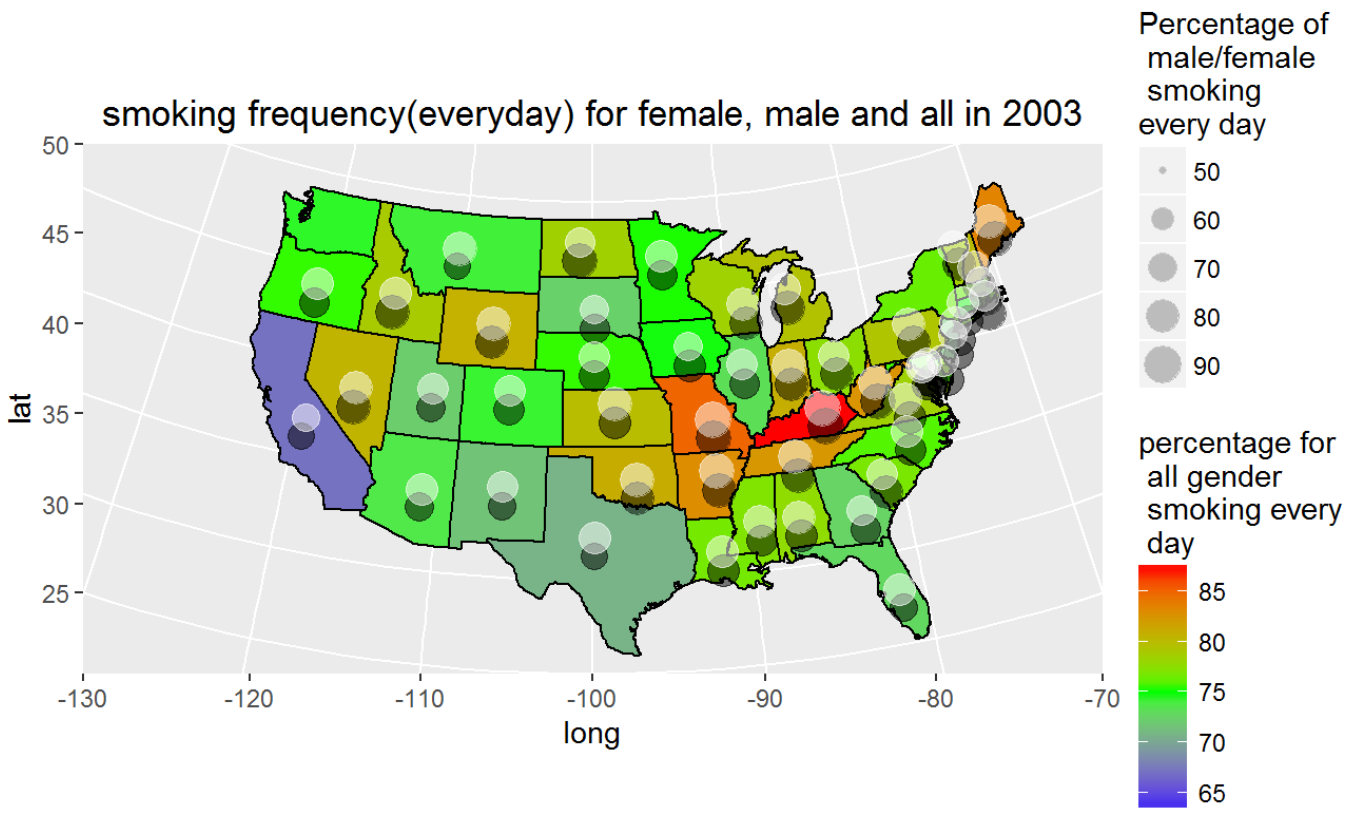
```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```

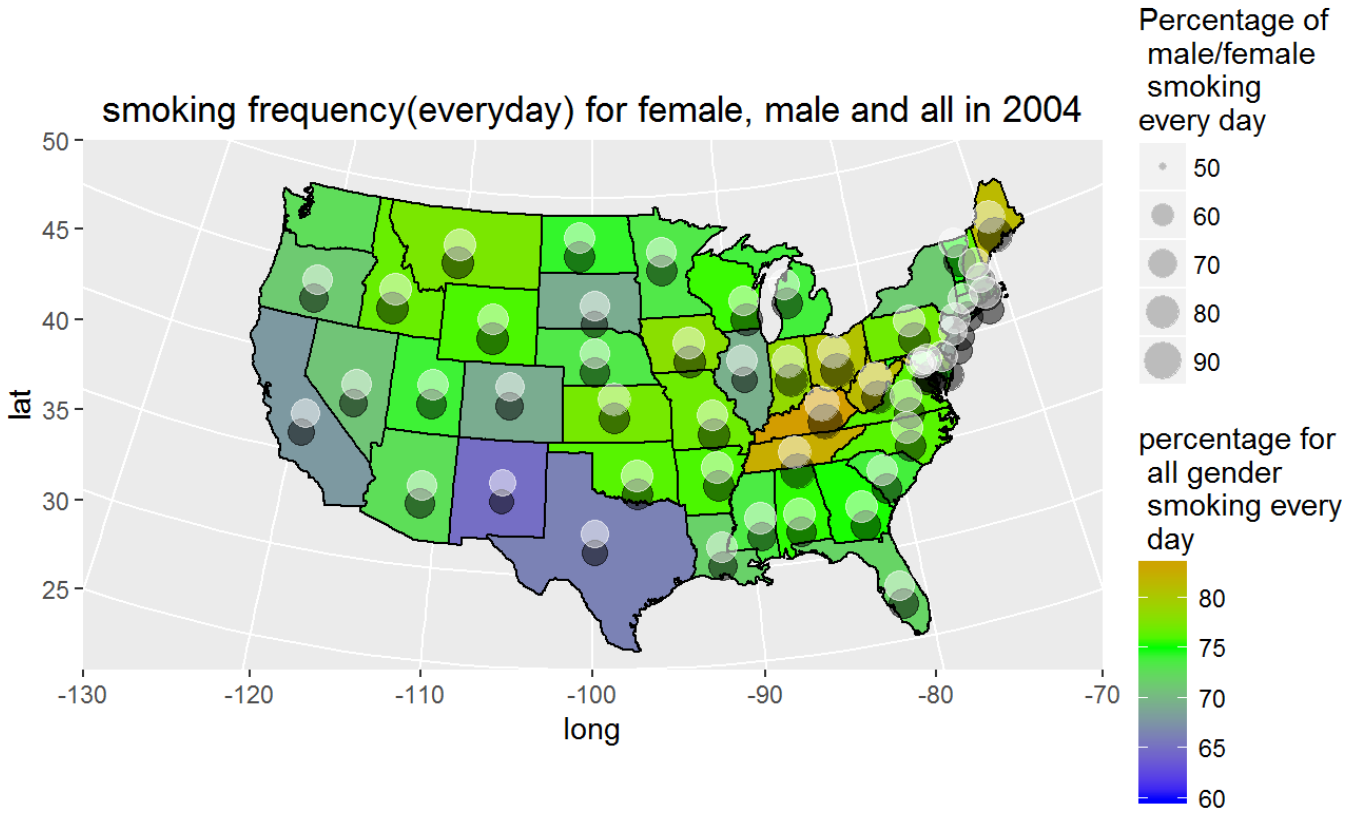



```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```



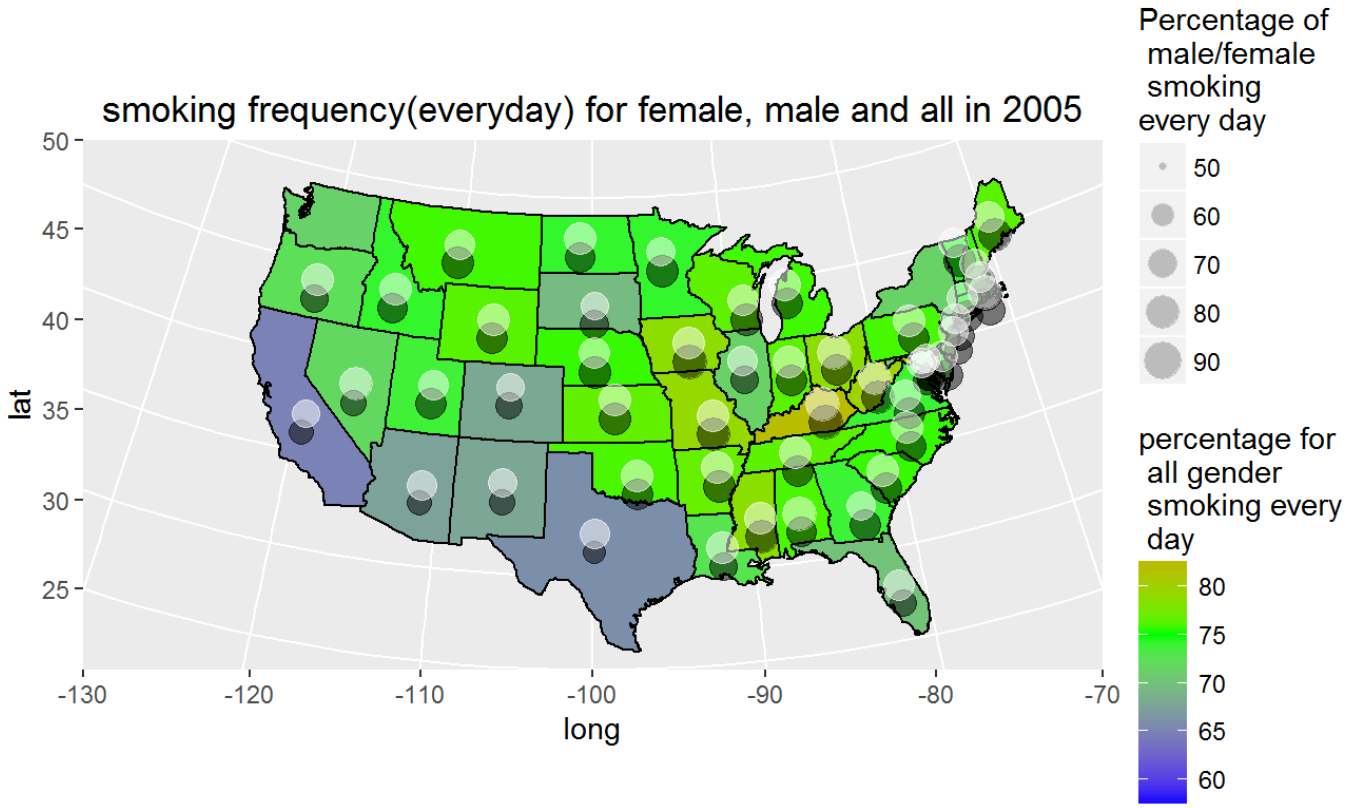

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

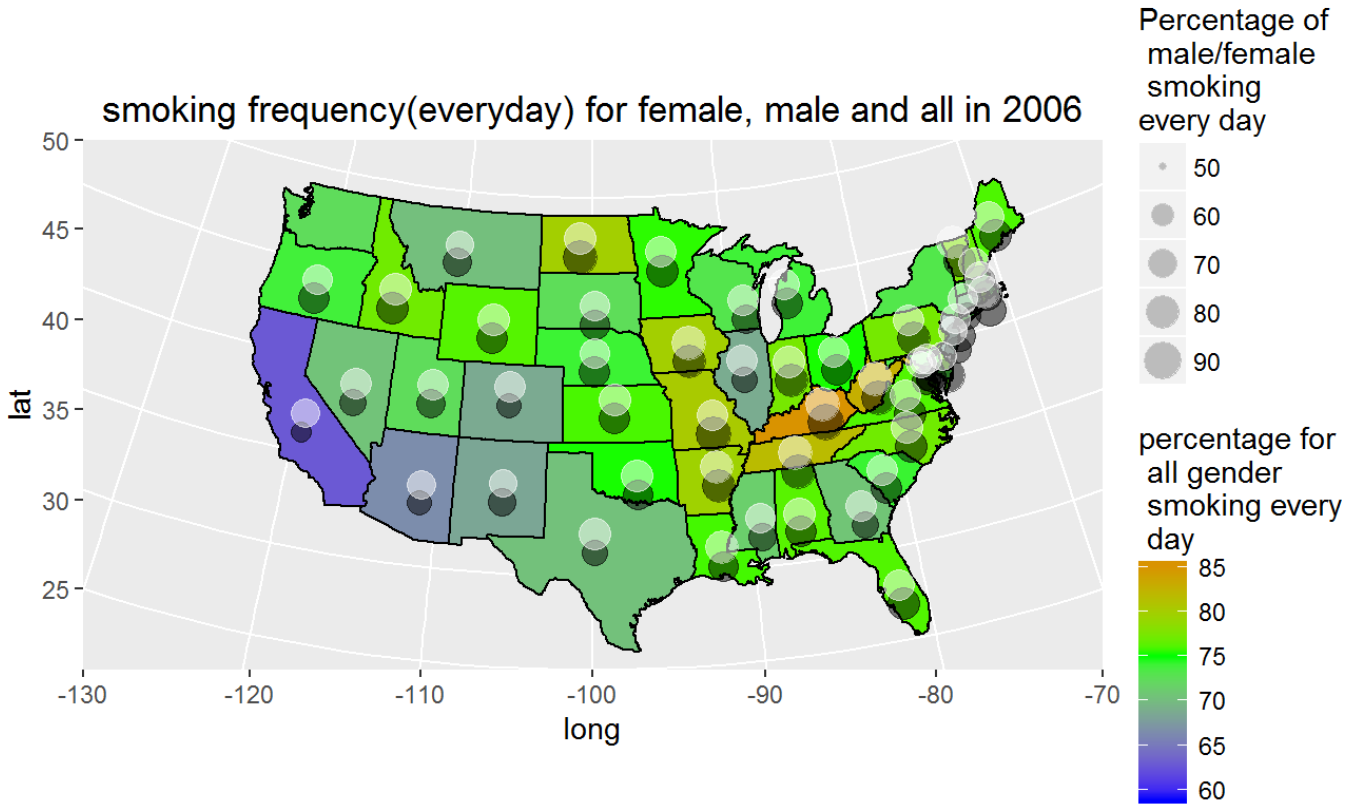



```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

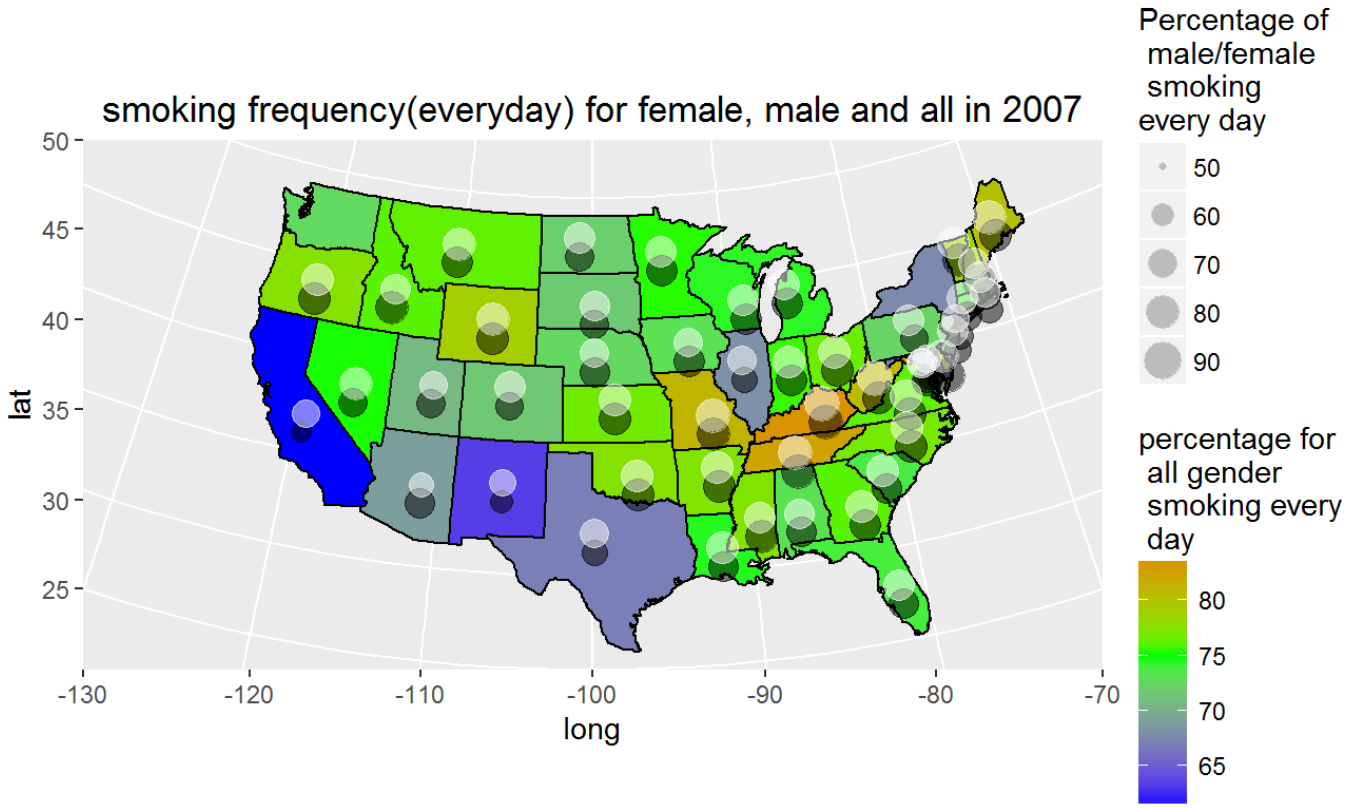



```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).
```

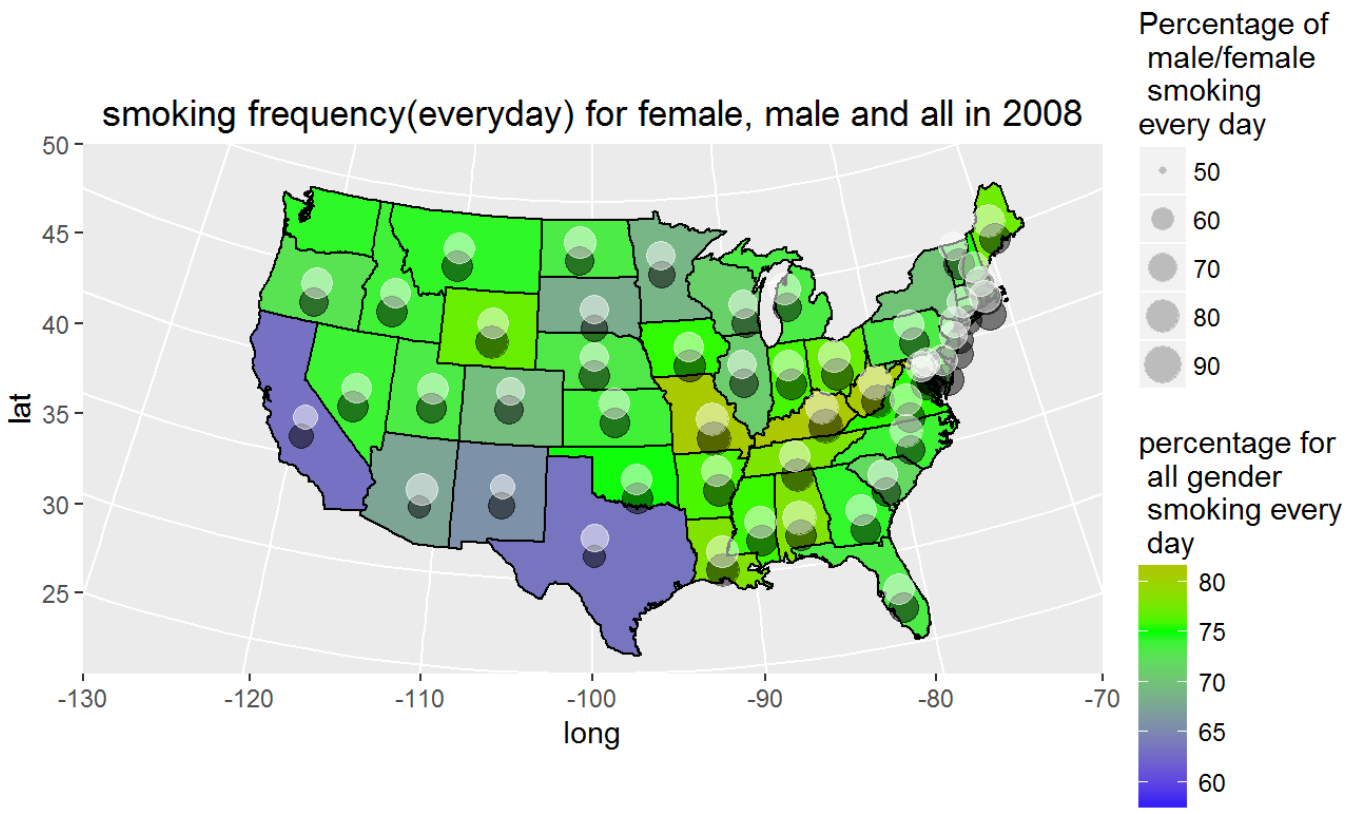



```
## Warning: Removed 1 rows containing missing values (geom_point).
```

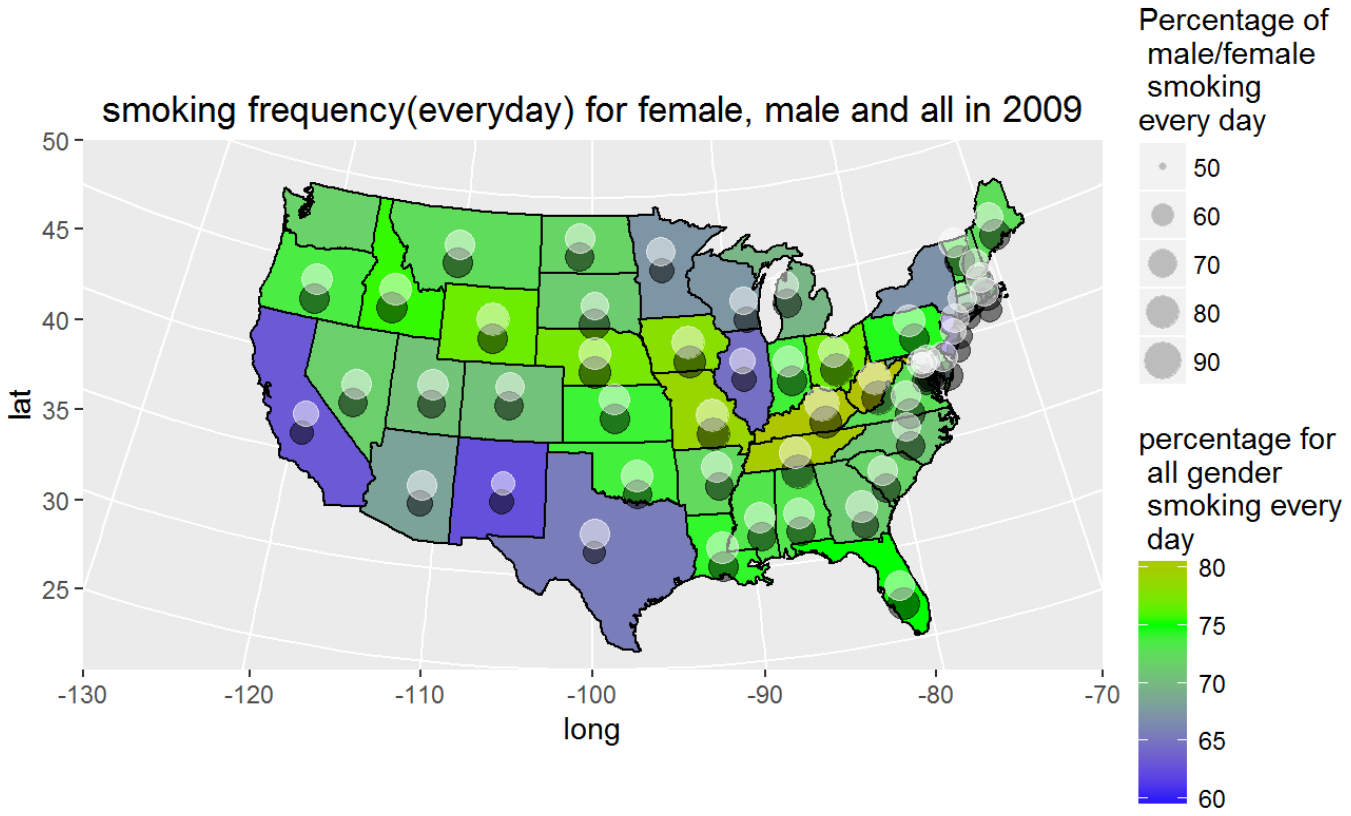
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



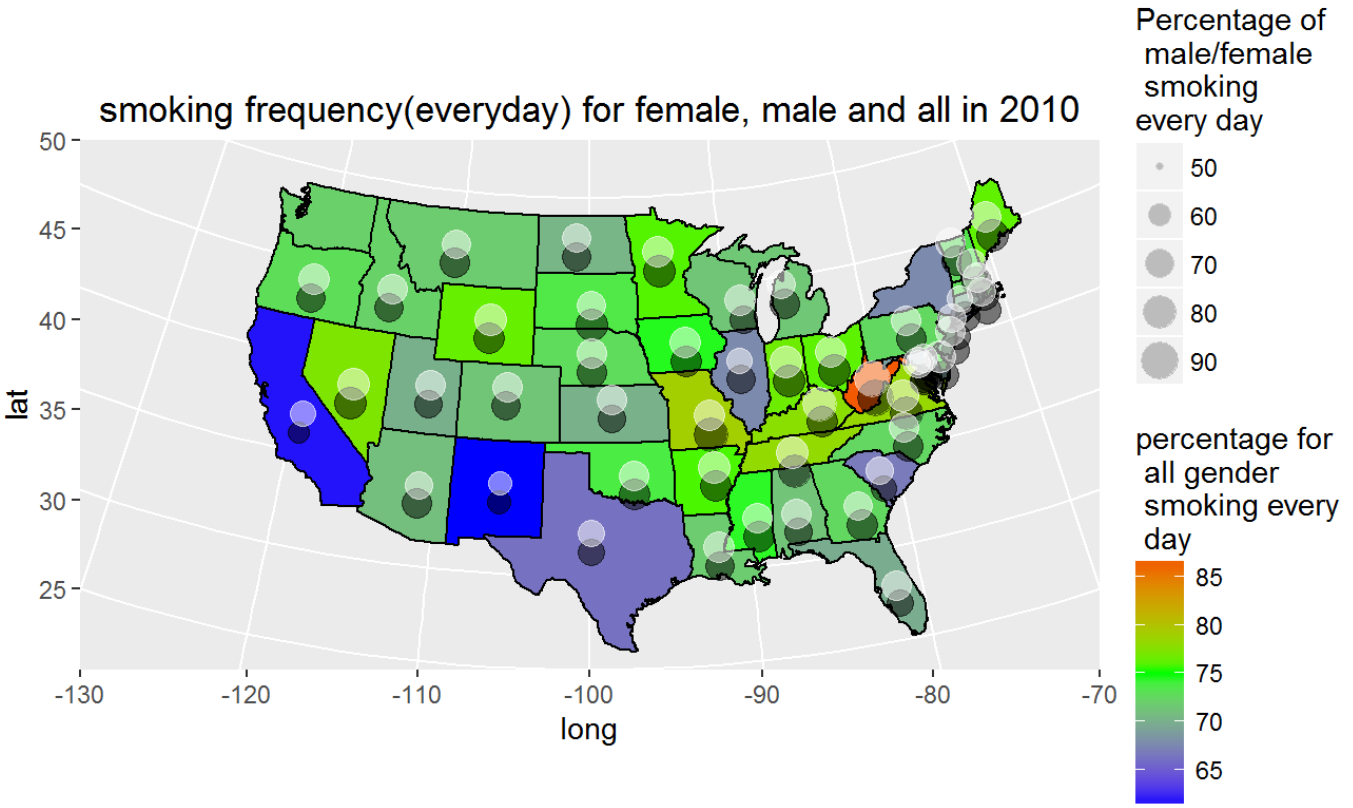

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```



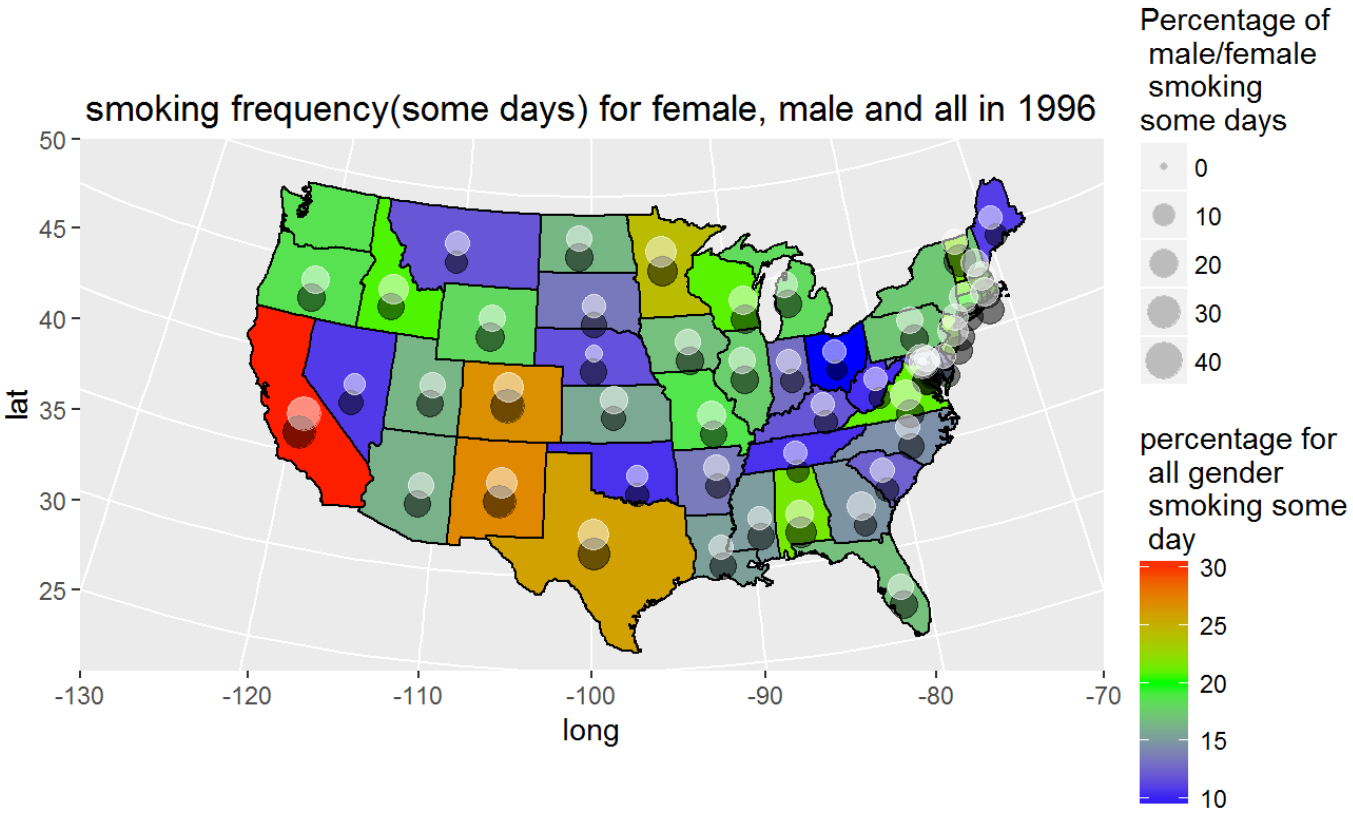

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```



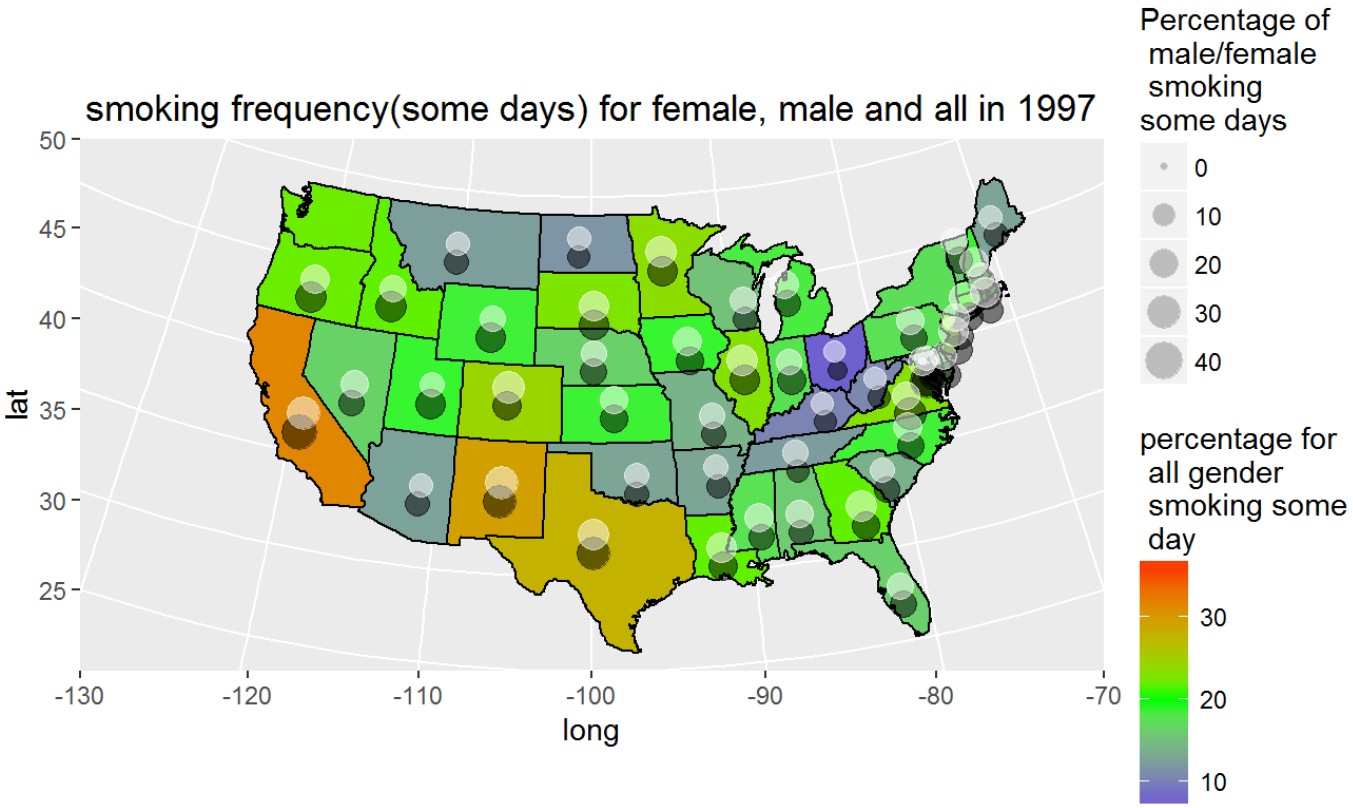

```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```

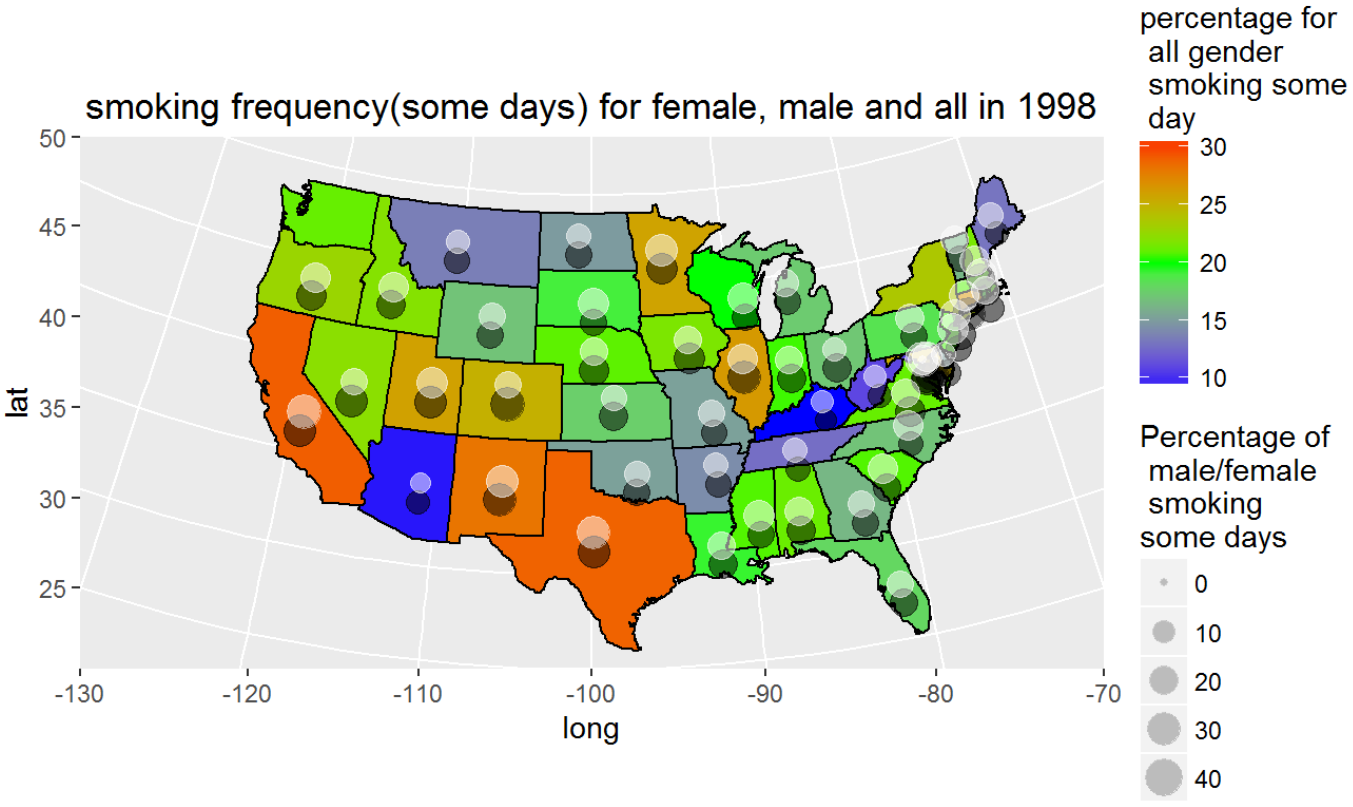


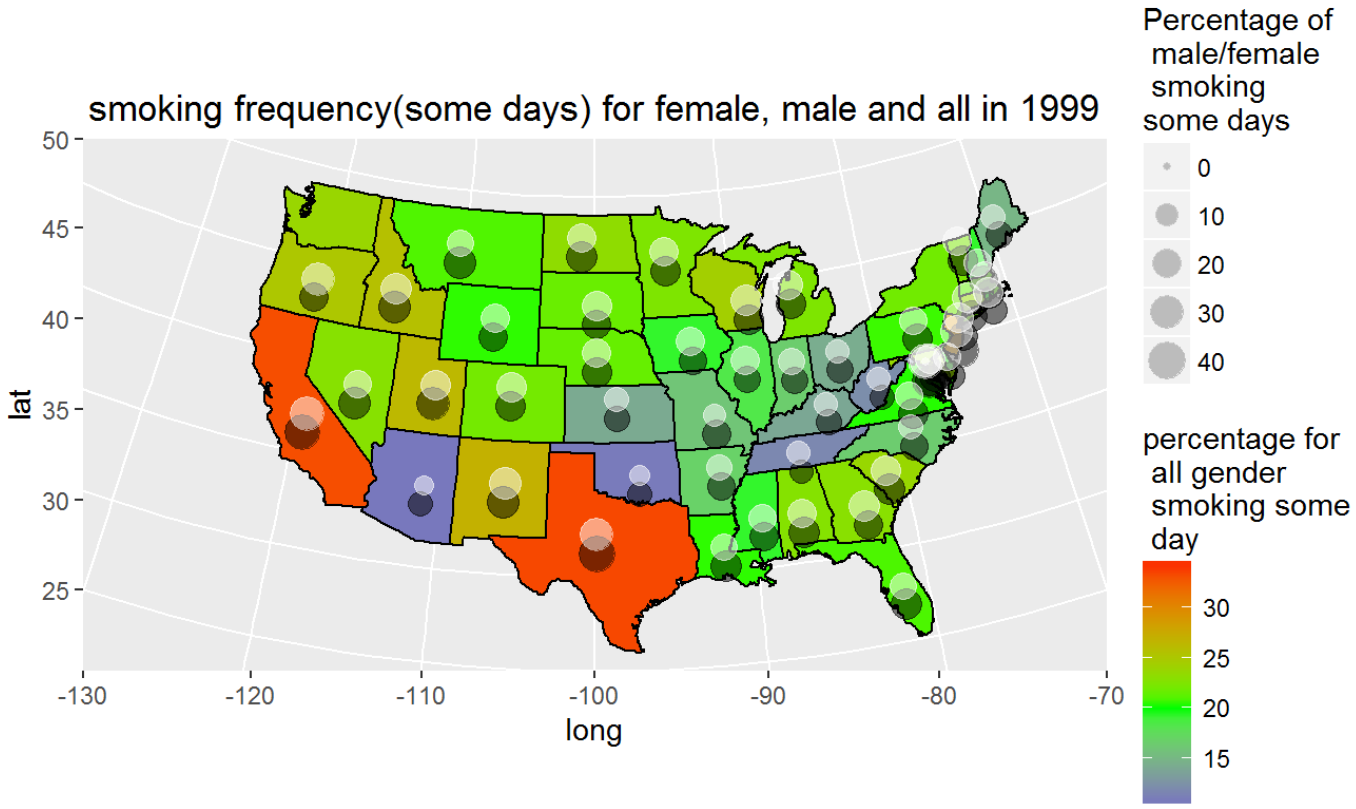
The plots above discribe the trend of people smoking every day changing in 15 years. We can tell that the overall percentage of smoking everyday decreases and percentage of people smoking every day does not distribute evenly among different states. The west coast has a lower percentage of people smoking every day than the east coast, the percentage of the middle part of the continential US is always between the one of the east coast and the one of the west coast. States “WV”, “AZ” and “KY” always has a higher percentage of people smoking everyday. There is no such a big difference between the percentage of male smoking everyday and the one of female smoking everyday but sometimes, female has a slightly larger circle than that of male, whcih means that female has a slightly higher percentage of smoking every day than male does.

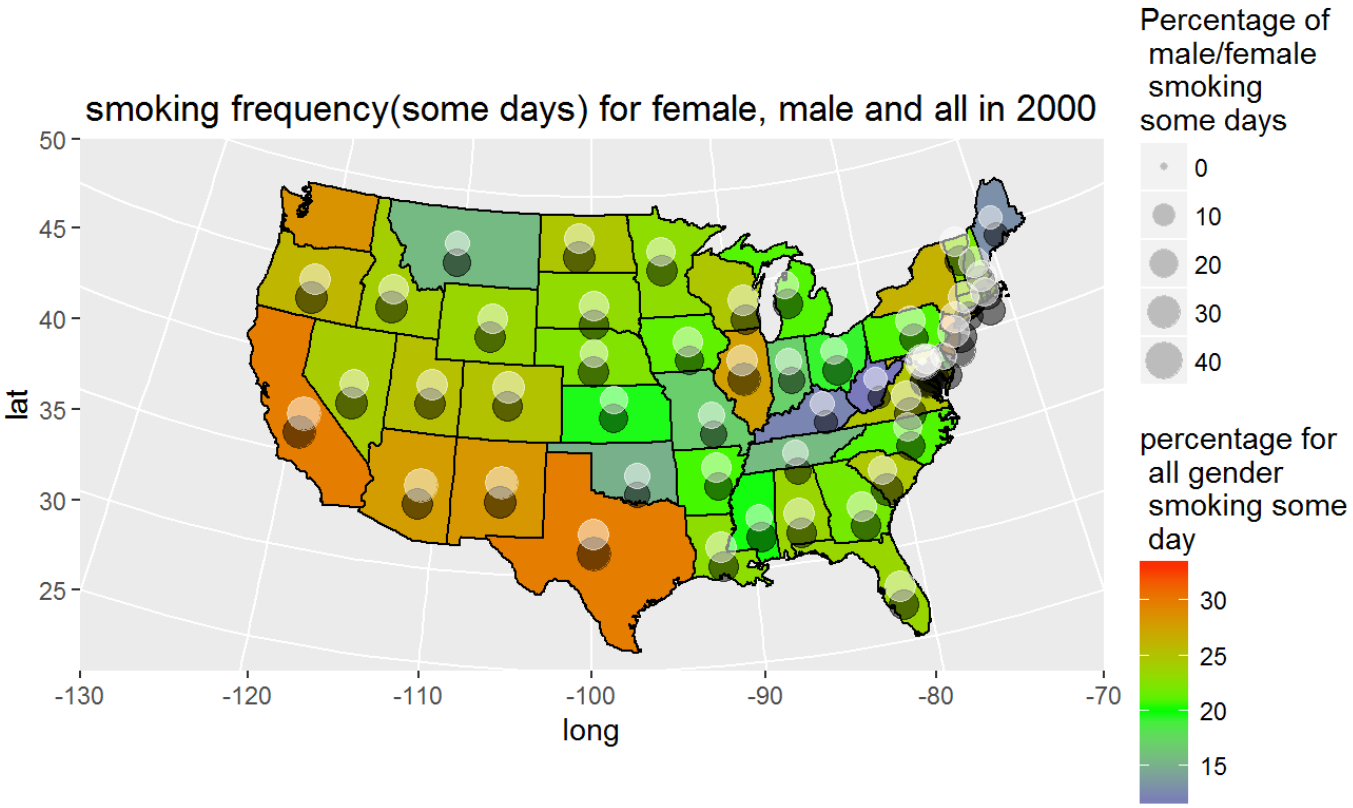



```
## Warning: Removed 1 rows containing missing values (geom_point).
```

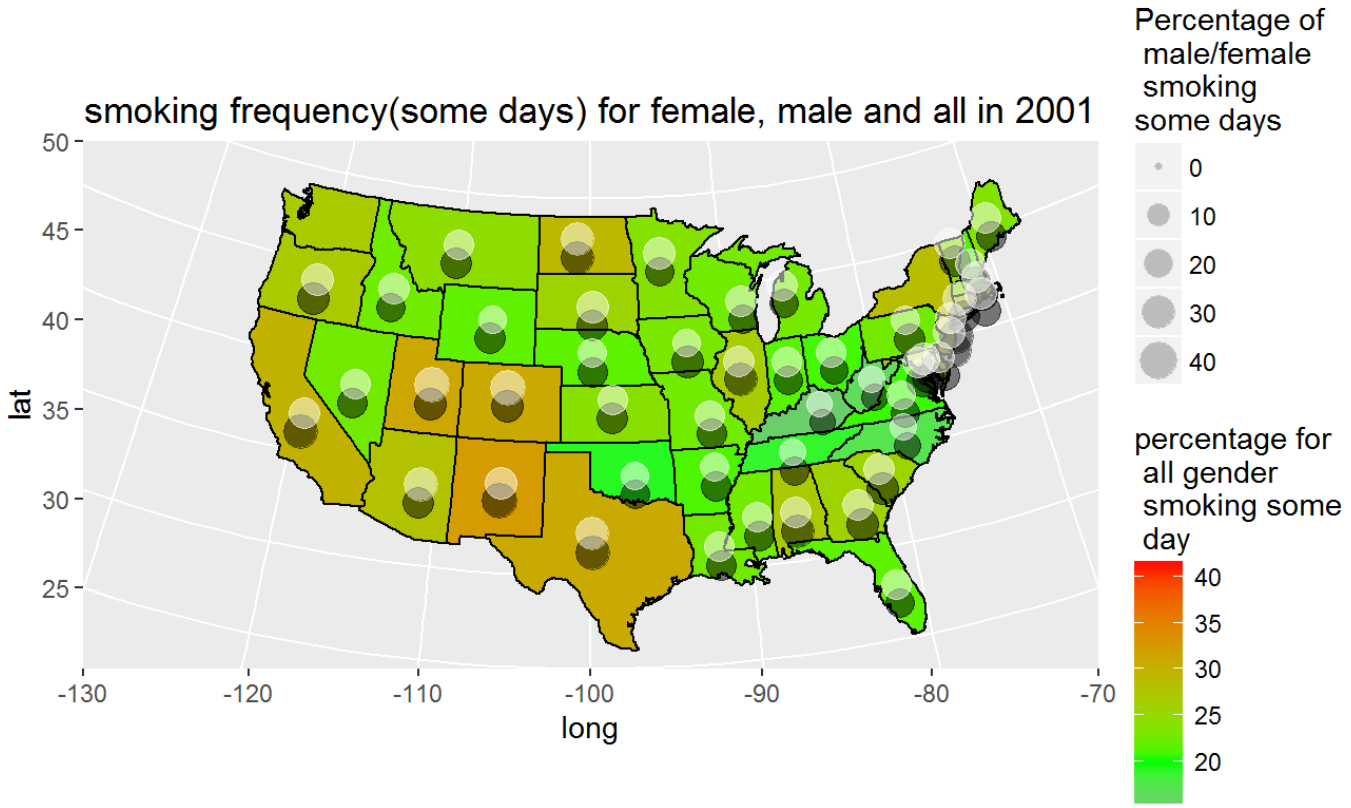

smoking frequency(some days) for female, male and all in 1997
lat
long
50
45
40
35
30
25
-130
-120
-110
-100
-90
-80
-70
Percentage of male/female smoking some days
0
10
20
30
40
percentage for all gender smoking some day
30
20
10

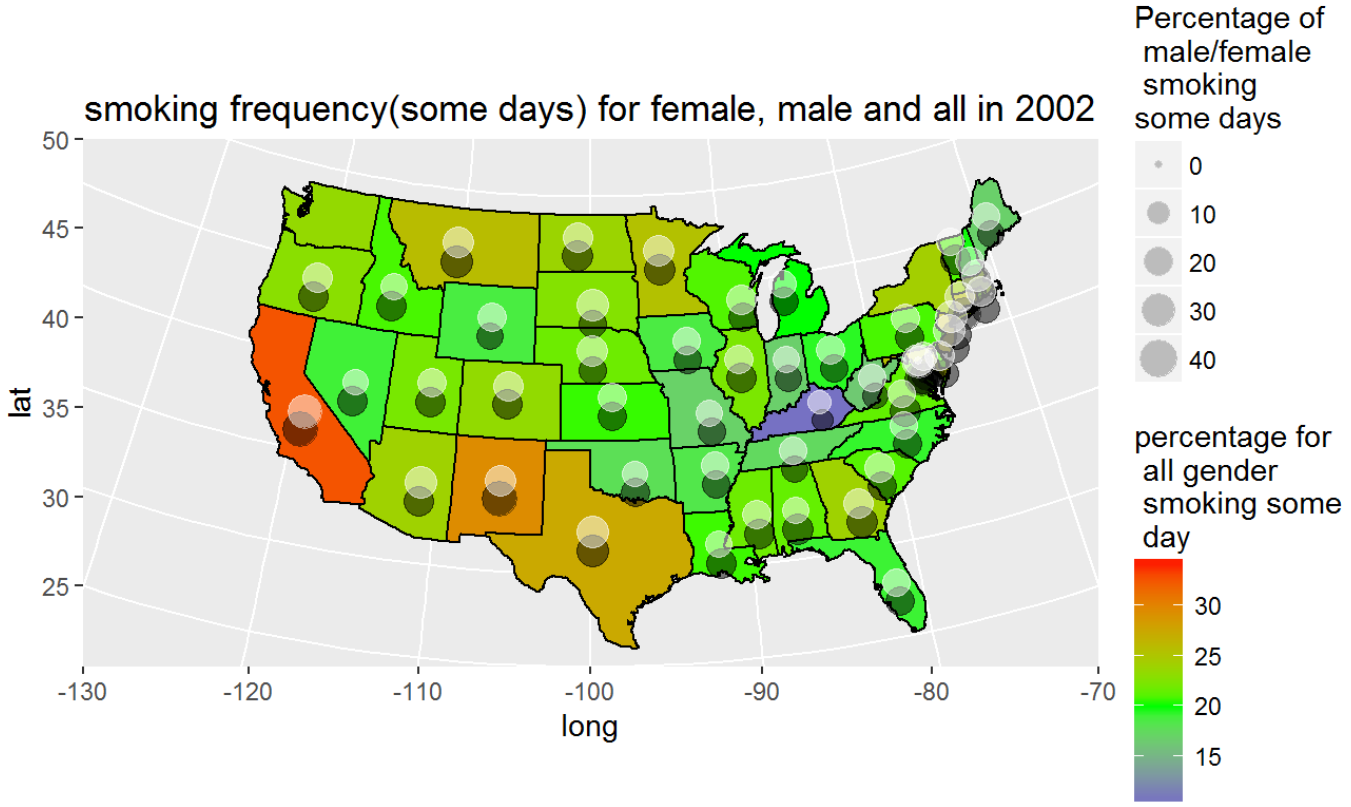

smoking frequency(some days) for female, male and all in 1998
lat
long
50
45
40
35
30
25
-130
-120
-110
-100
-90
-80
-70
percentage for all gender smoking some day
30
25
20
15
10
Percentage of male/female smoking some days
0
10
20
30
40

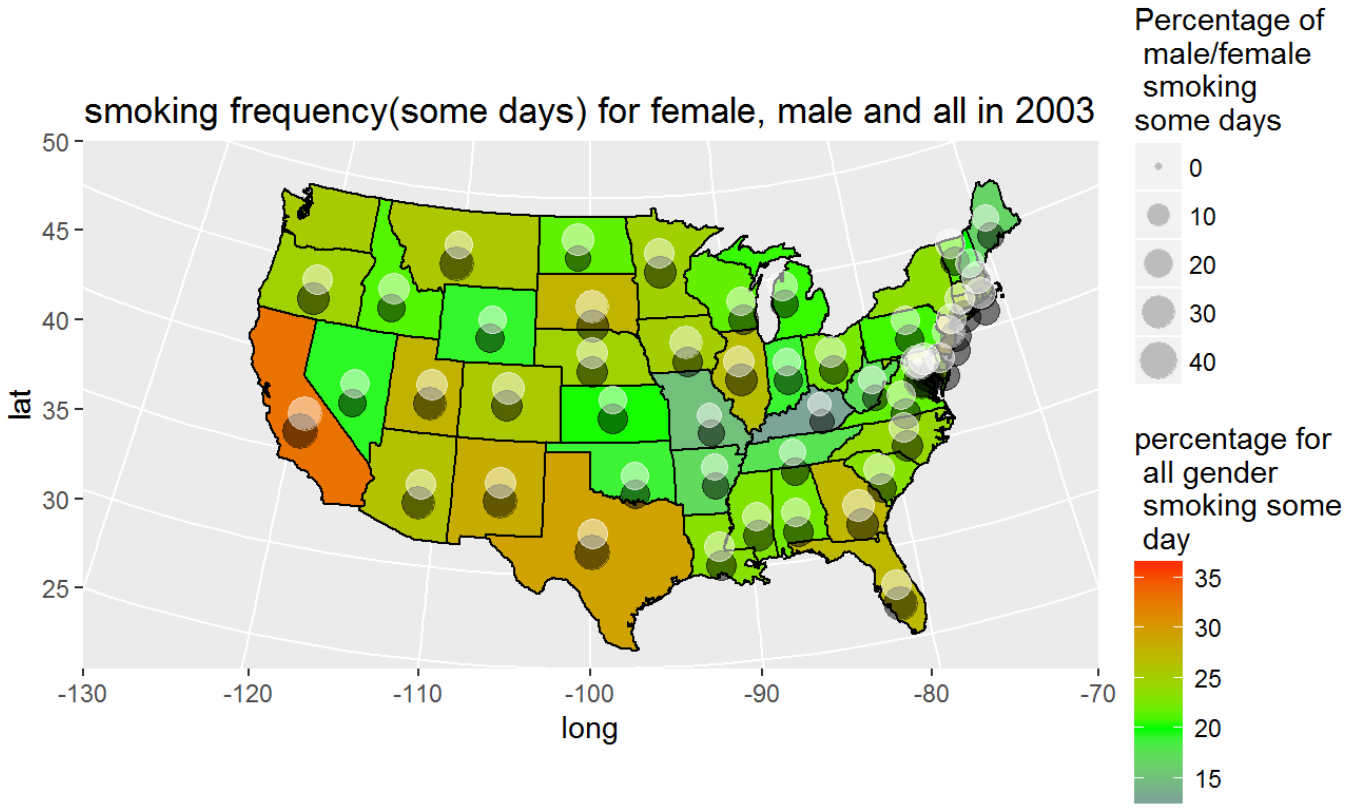

smoking frequency(some days) for female, male and all in 1999
lat
long
50
45
40
35
30
25
-130
-120
-110
-100
-90
-80
-70
Percentage of male/female smoking some days
0
10
20
30
40
percentage for all gender smoking some day
30
25
20
15



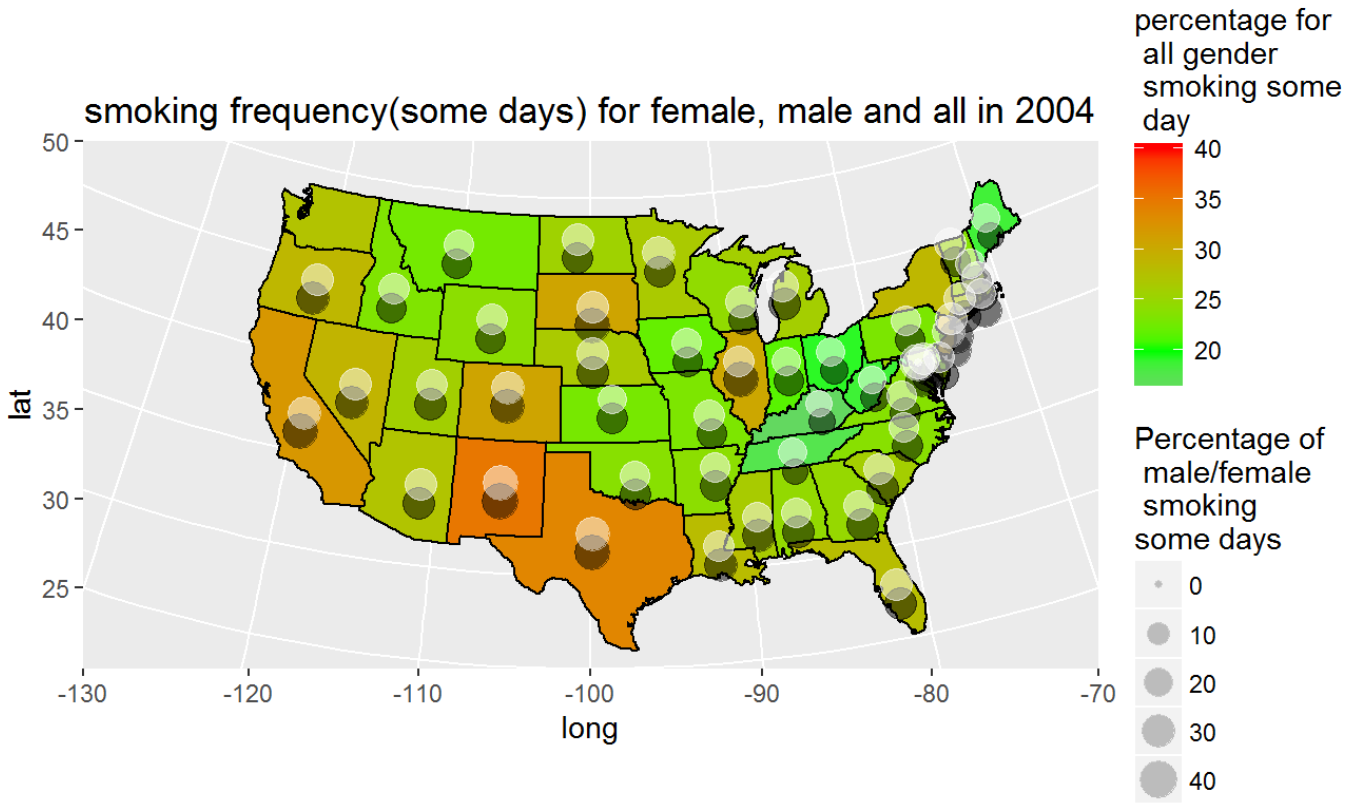

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```


smoking frequency(some days) for female, male and all in 2001
lat
long
50
45
40
35
30
25
-130
-120
-110
-100
-90
-80
-70
Percentage of male/female smoking some days
0
10
20
30
40
percentage for all gender smoking some day
40
35
30
25
20

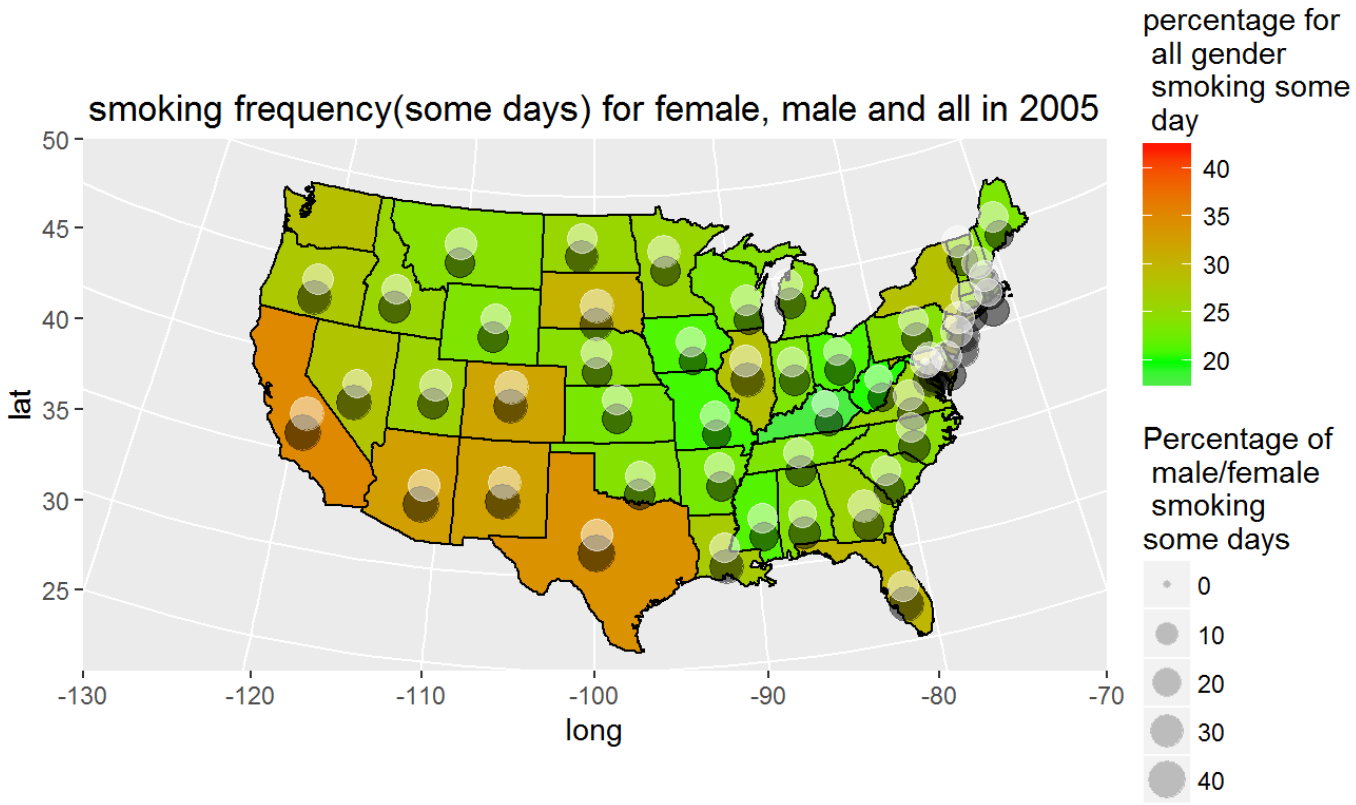

smoking frequency(some days) for female, male and all in 2002
lat
long
50
45
40
35
30
25
-130
-120
-110
-100
-90
-80
-70
Percentage of male/female smoking some days
0
10
20
30
40
percentage for all gender smoking some day
30
25
20
15



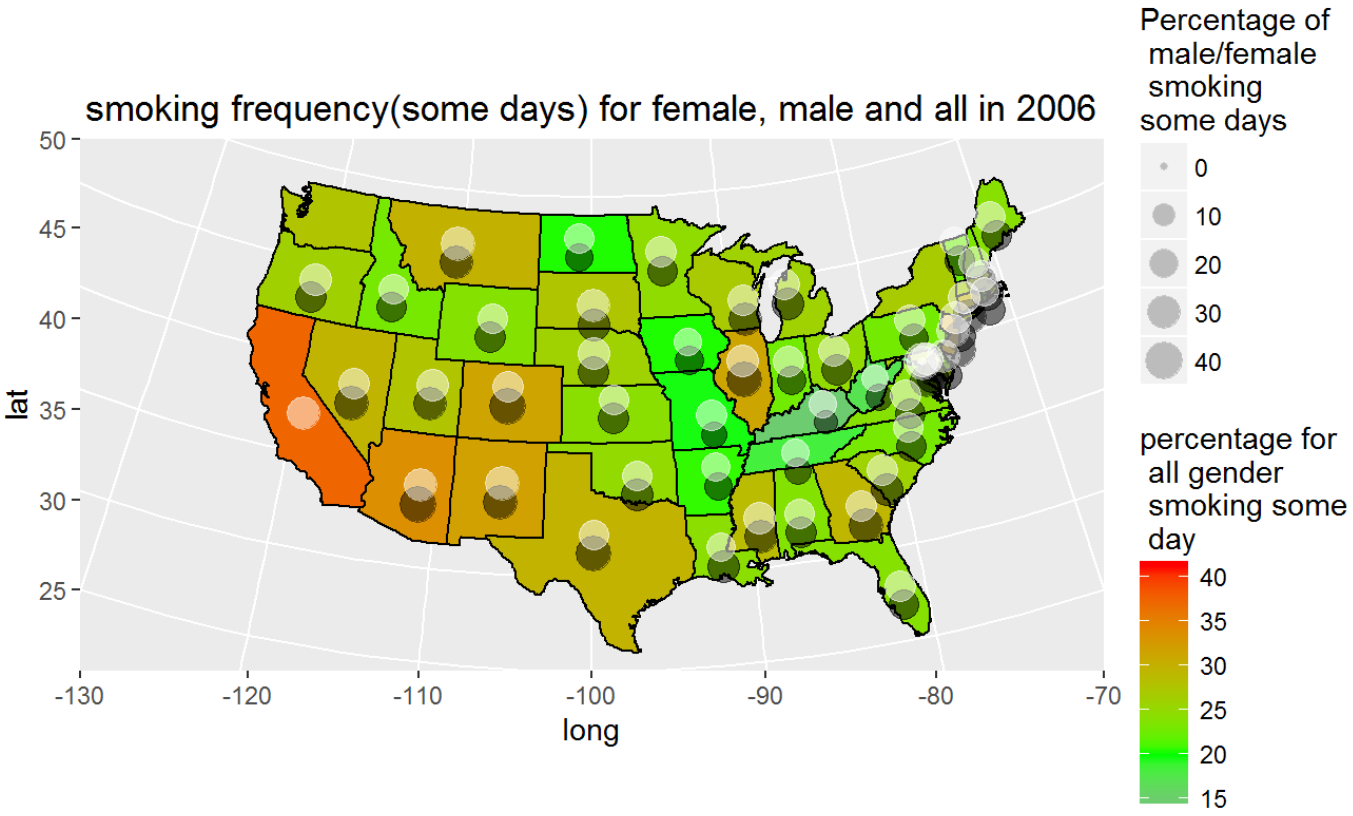

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



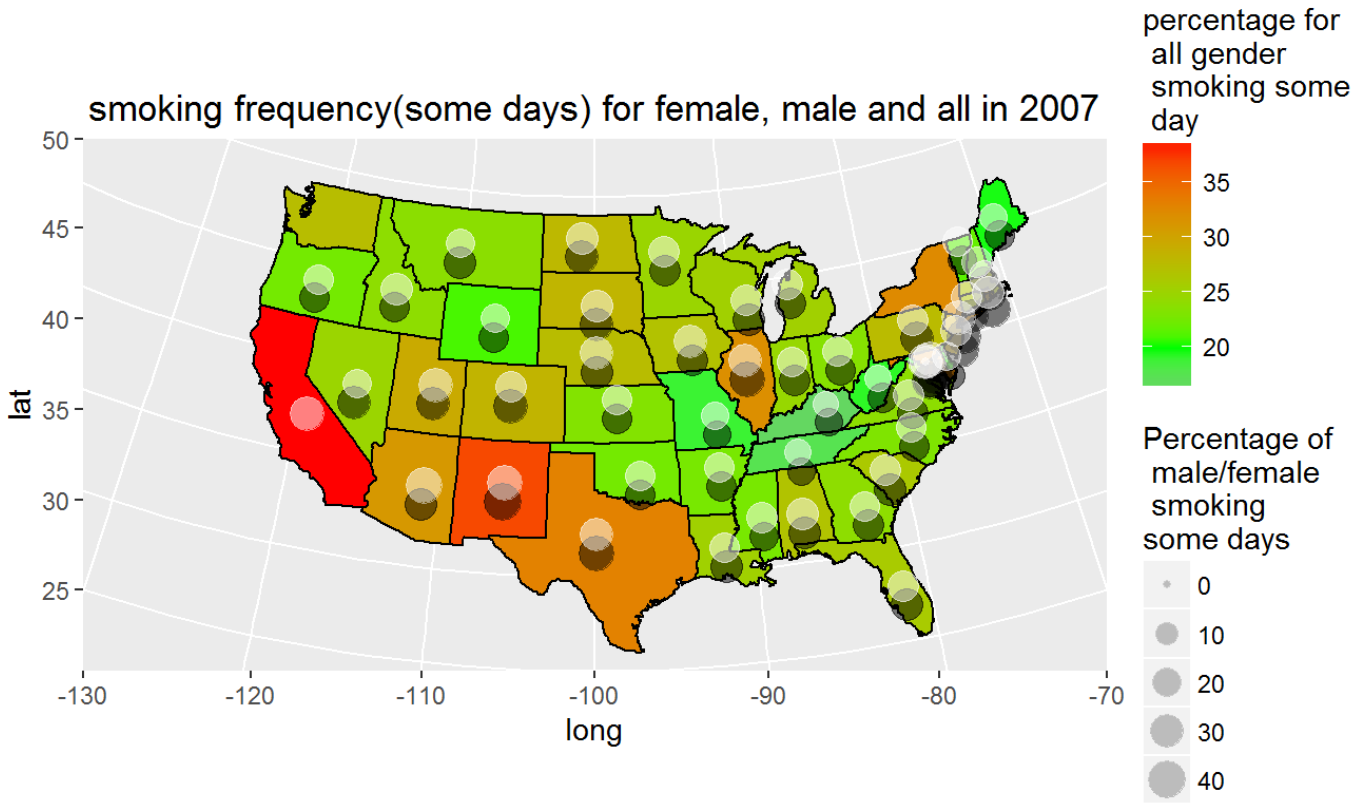

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```




```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```

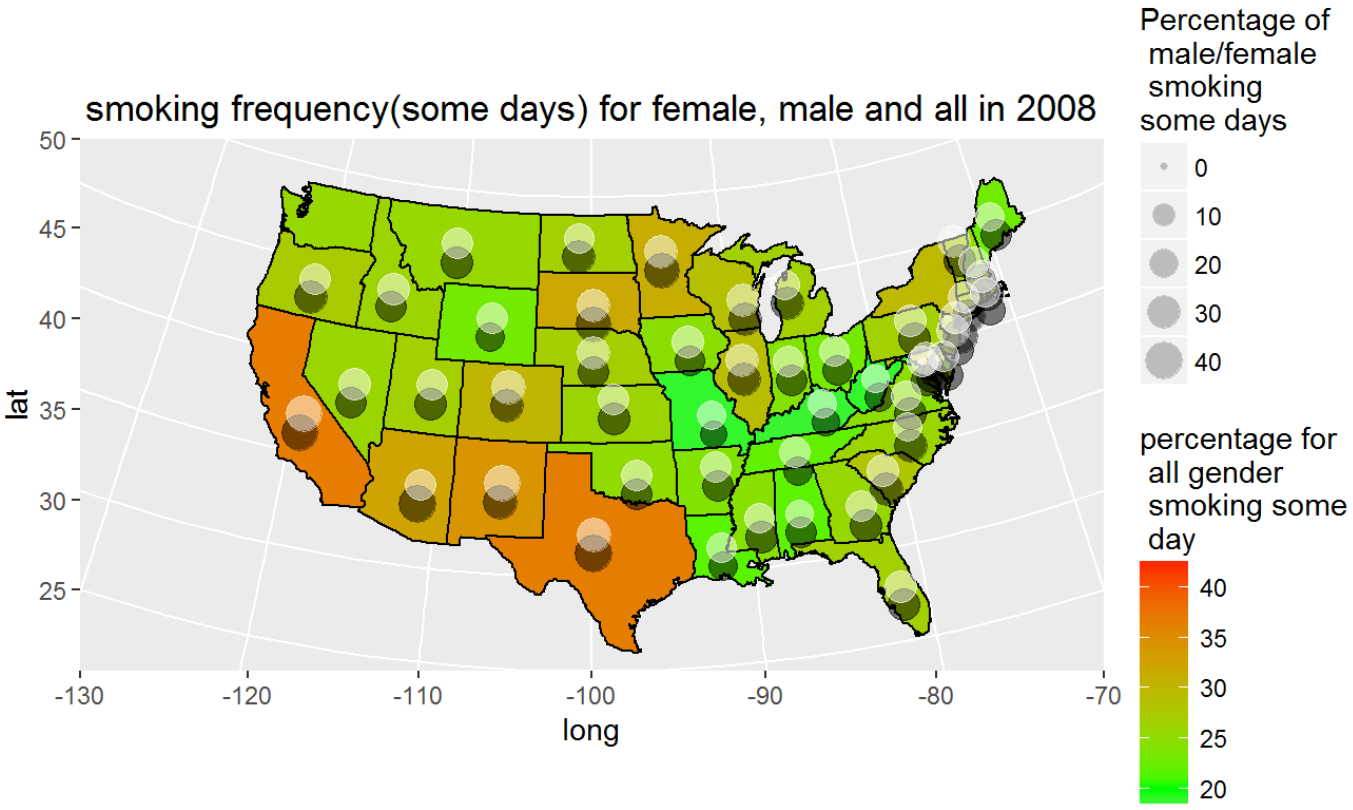



```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).
```

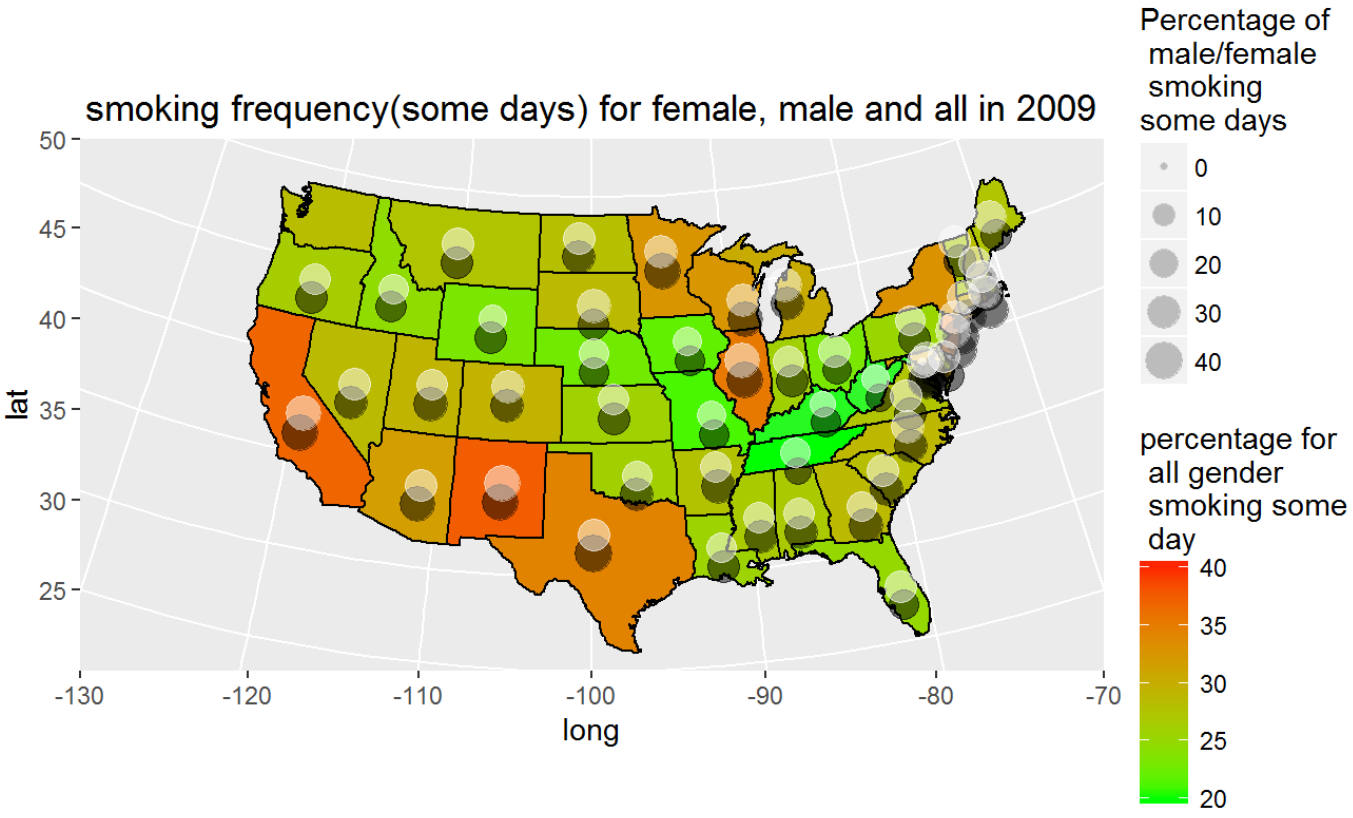



```
## Warning: Removed 1 rows containing missing values (geom_point).
```

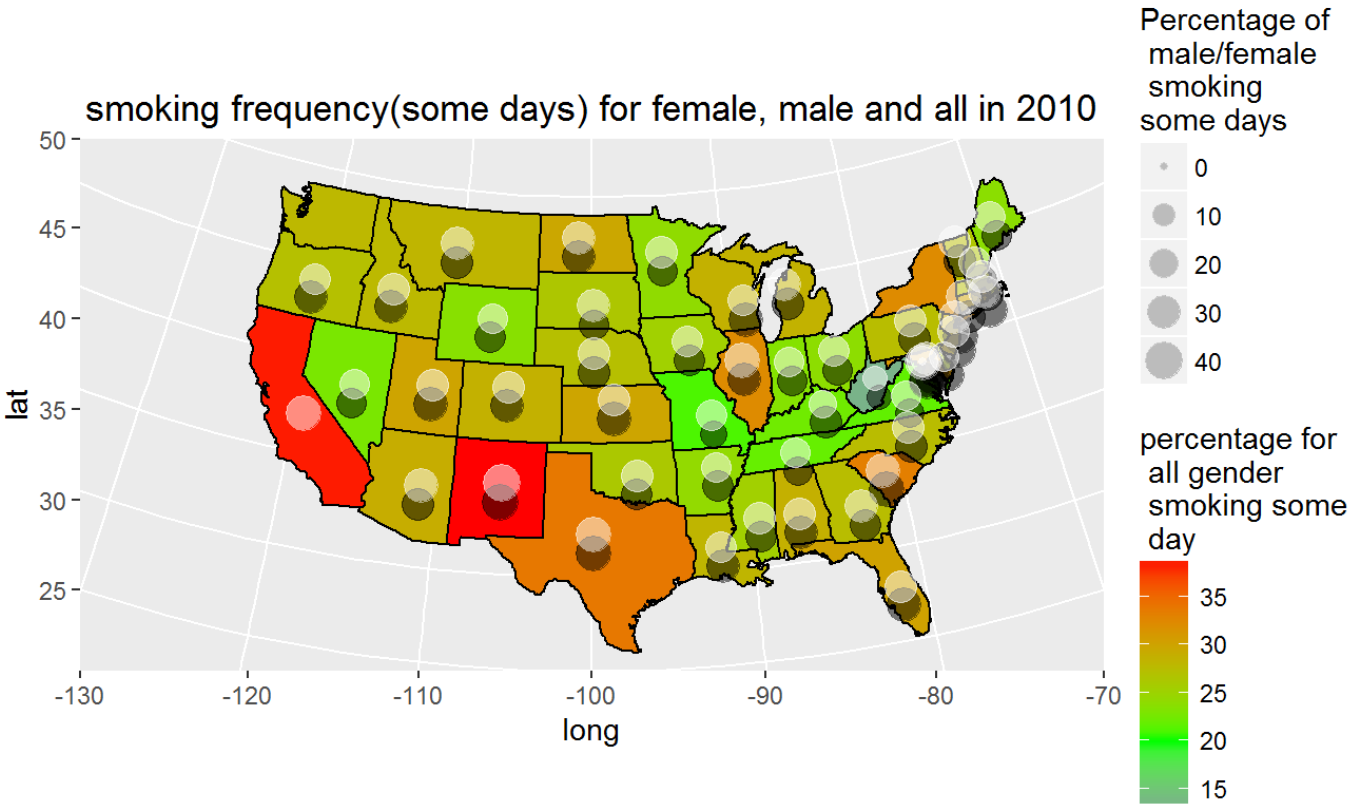
```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).
```

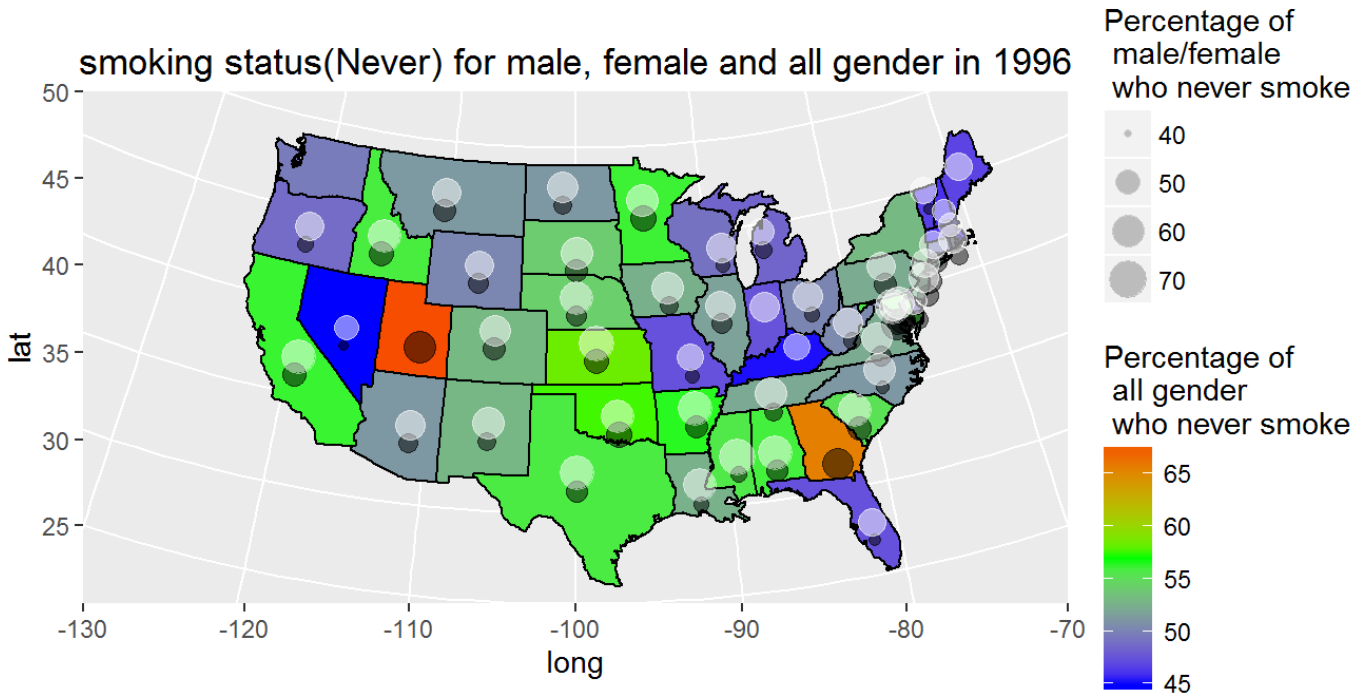



```
## Warning: Removed 1 rows containing missing values (geom_point).
```



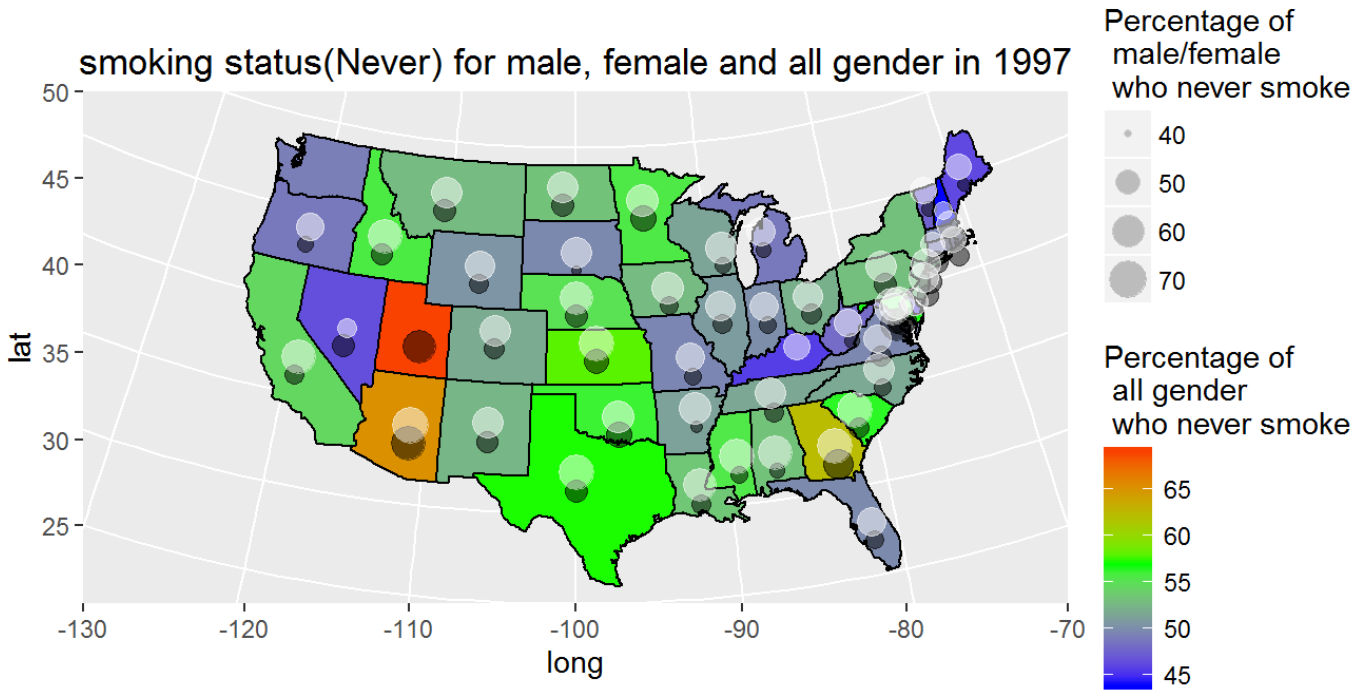
The plots above discribe the trend of people smoking some days changing in 15 years. We can tell that the overall percentage of smoking some days increases and the percentage of people smoking some days does not distribute evenly among different states. It is hard to tell the distribution in the first year, but in the second year, 1997, states with higher percentage of people smoking some days decrease their percentage and states with lower percentage of people smoking some days increases their percentage, which means that the overall trend spreads more evenly than the first year. Same thing happened in 2001. States “CA”, “TX” and “NM” has a higher percentage of people smoking some days than the rest other states. It is hard to tell whether more percentage of female smoke more and that of male, because the percentages of female and male change all the time.

```
## Warning: Removed 6 rows containing missing values (geom_point).

## Warning: Removed 6 rows containing missing values (geom_point).
```

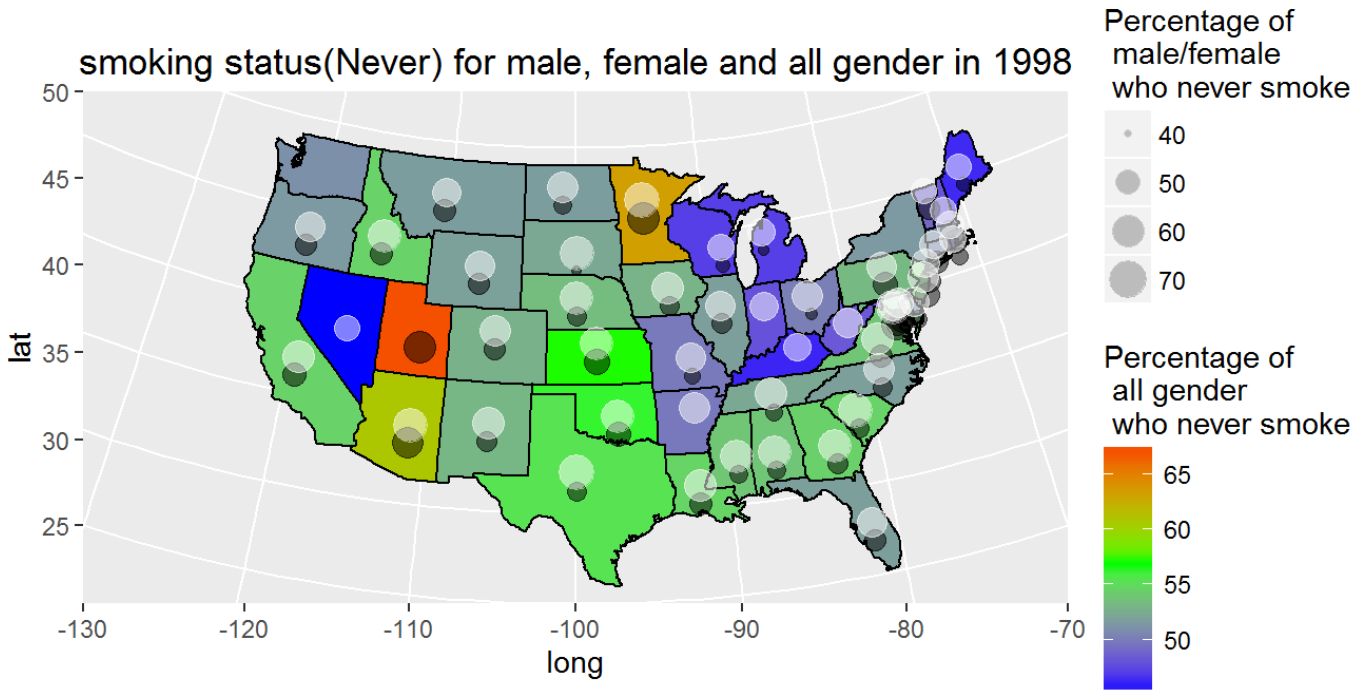



```
## Warning: Removed 5 rows containing missing values (geom_point).
```

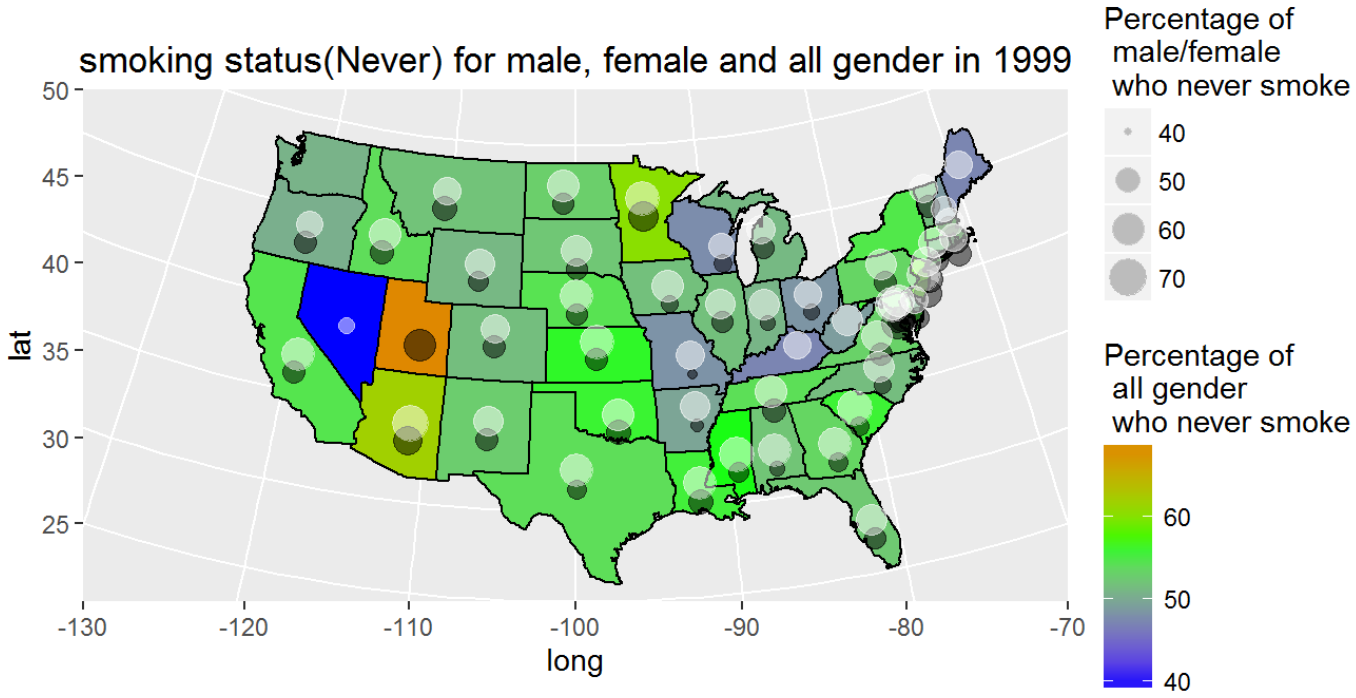
```
## Warning: Removed 5 rows containing missing values (geom_point).
```



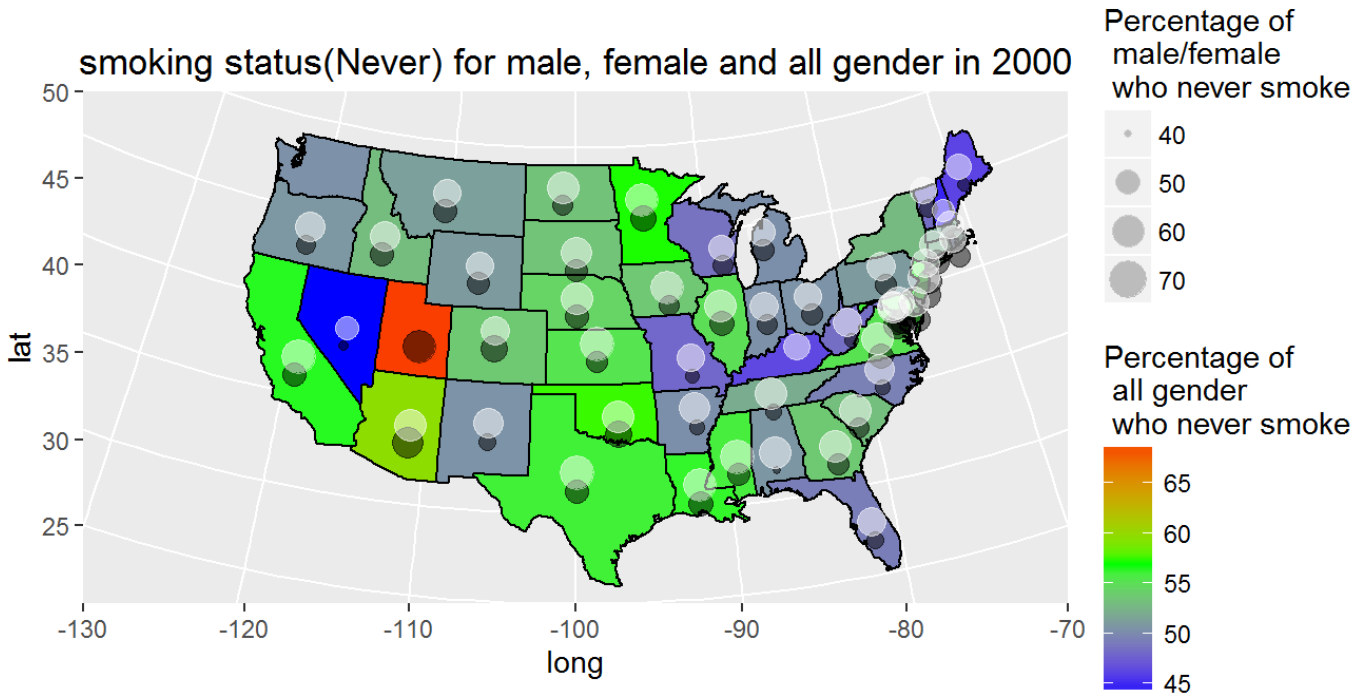

```
## Warning: Removed 8 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```




```
## Warning: Removed 7 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```

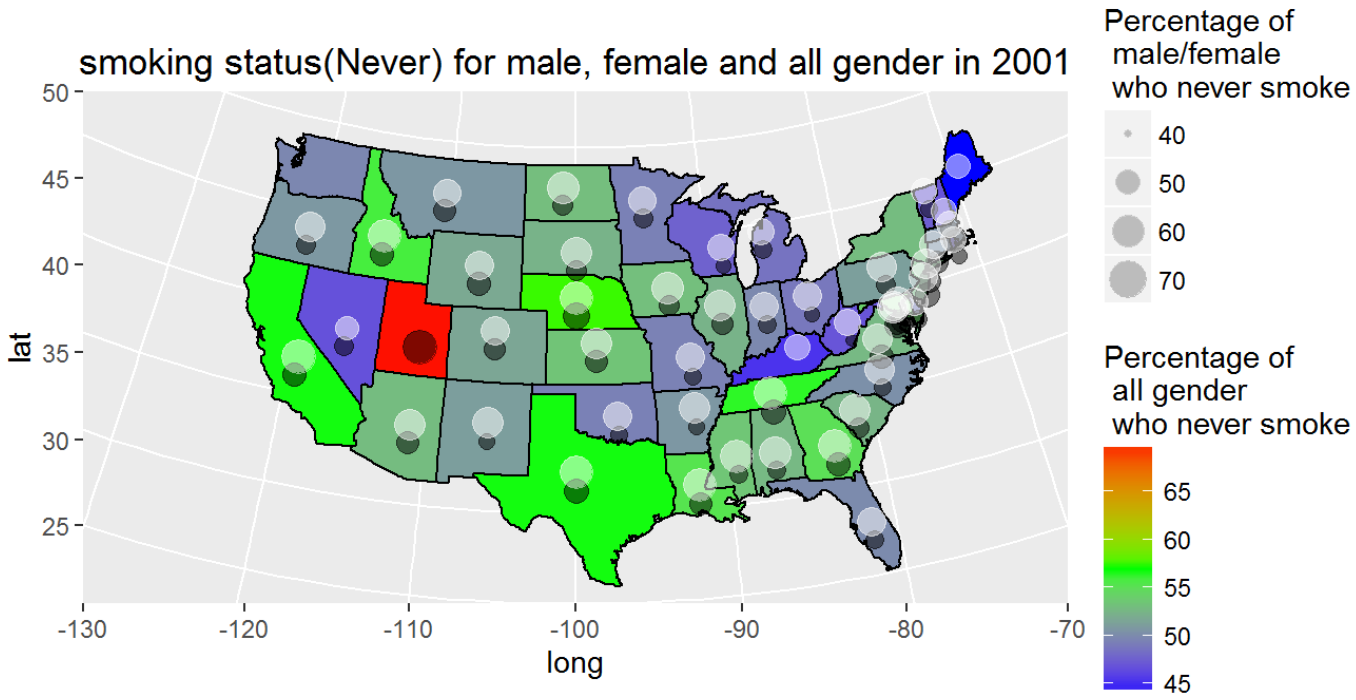



```
## Warning: Removed 4 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```

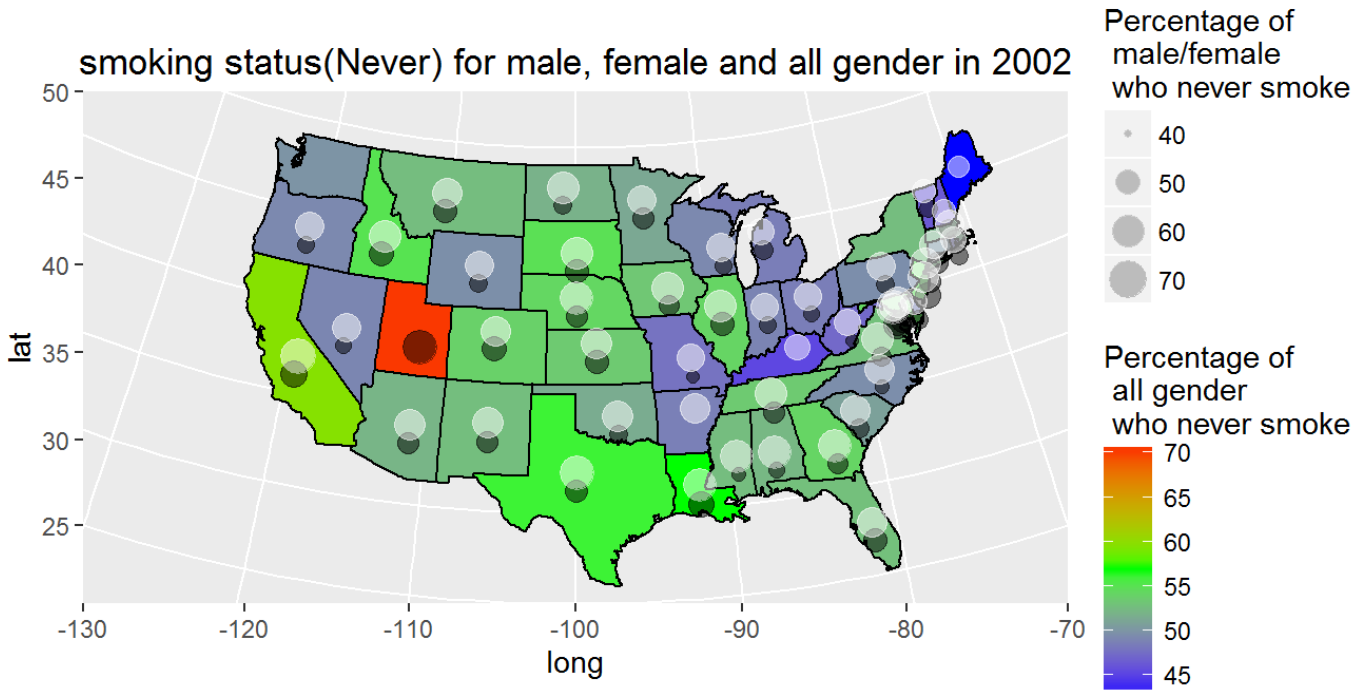



```
## Warning: Removed 3 rows containing missing values (geom_point).
```

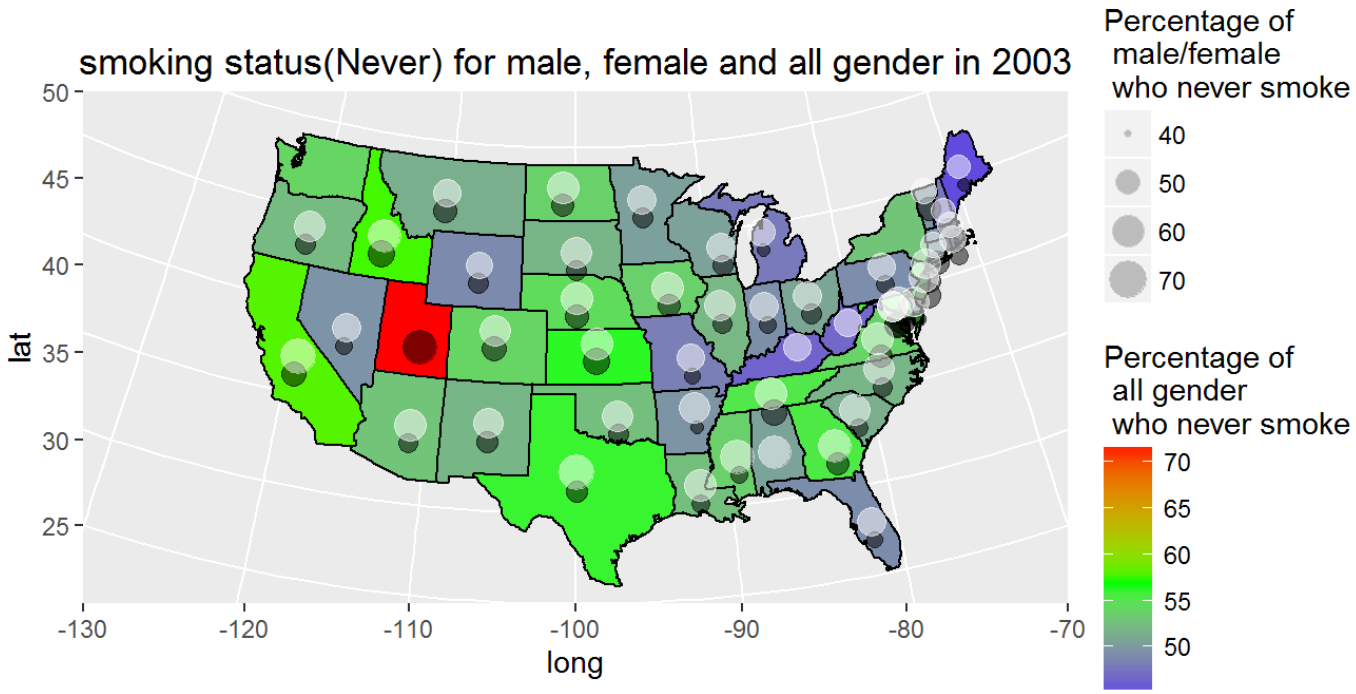
```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
## Warning: Removed 6 rows containing missing values (geom_point).

## Warning: Removed 4 rows containing missing values (geom_point).
```

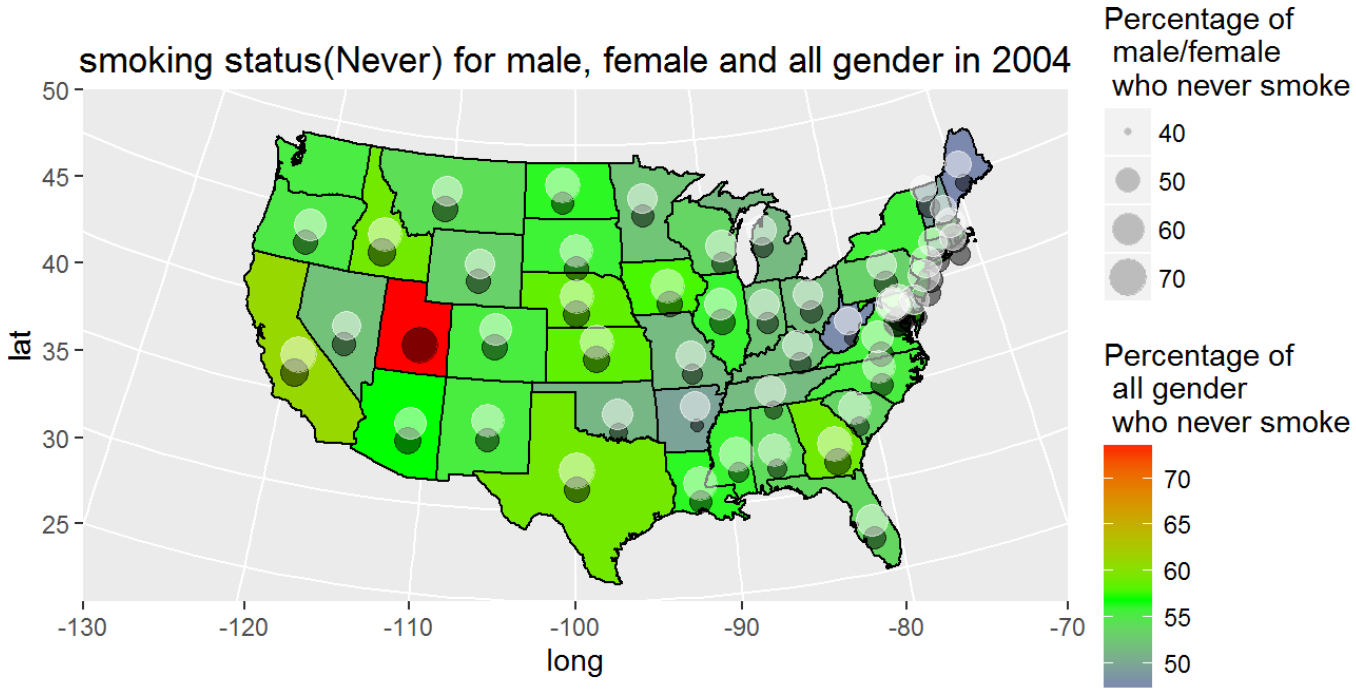



```
## Warning: Removed 5 rows containing missing values (geom_point).

## Warning: Removed 4 rows containing missing values (geom_point).
```



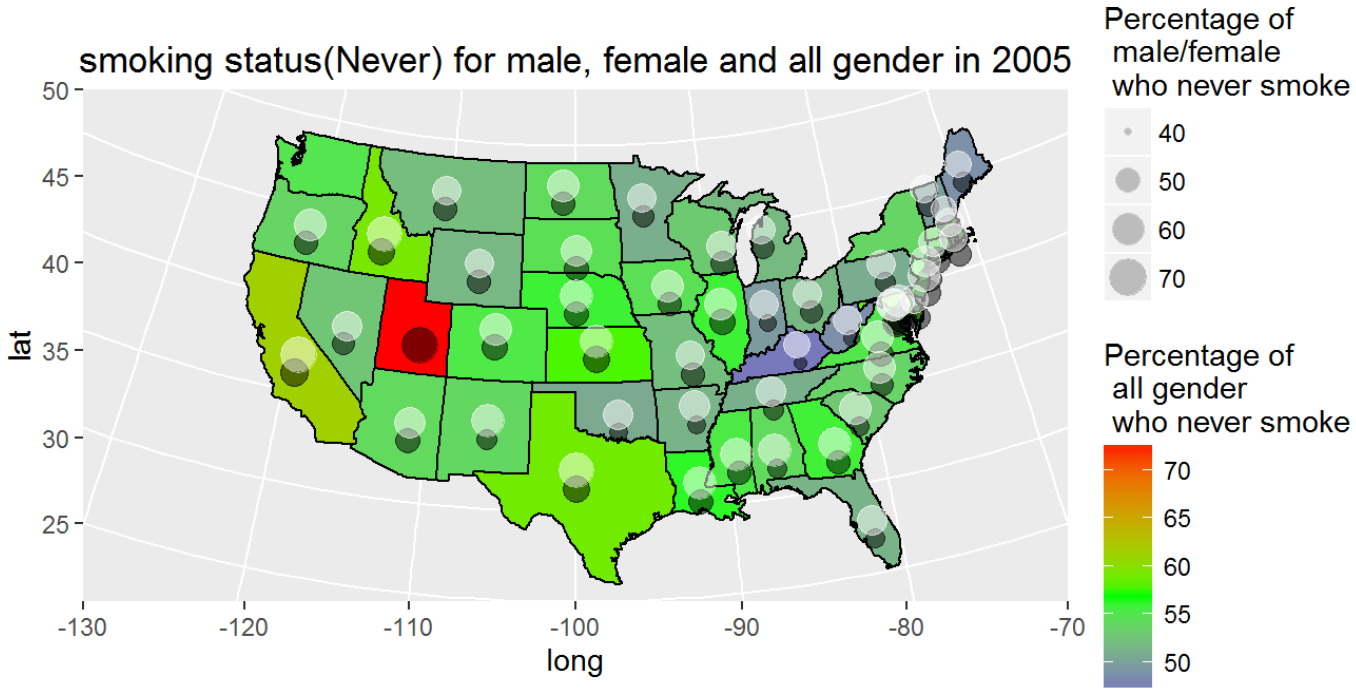

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```

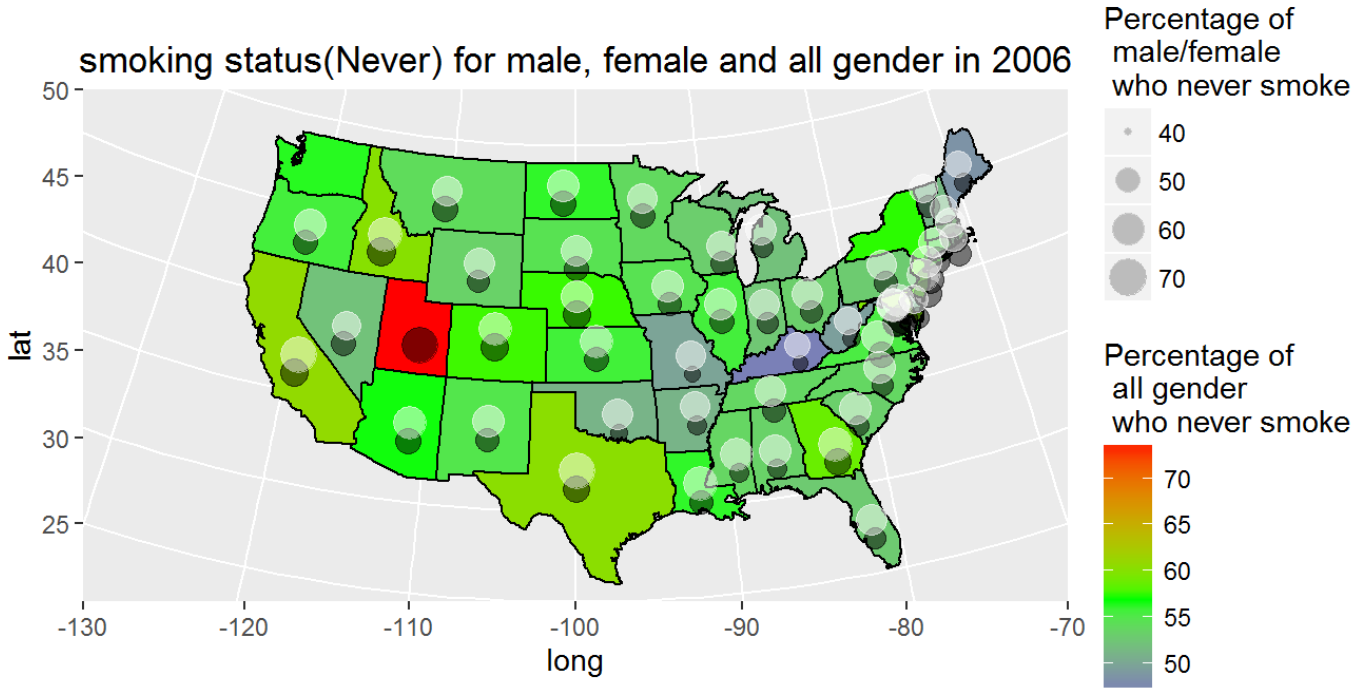



```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

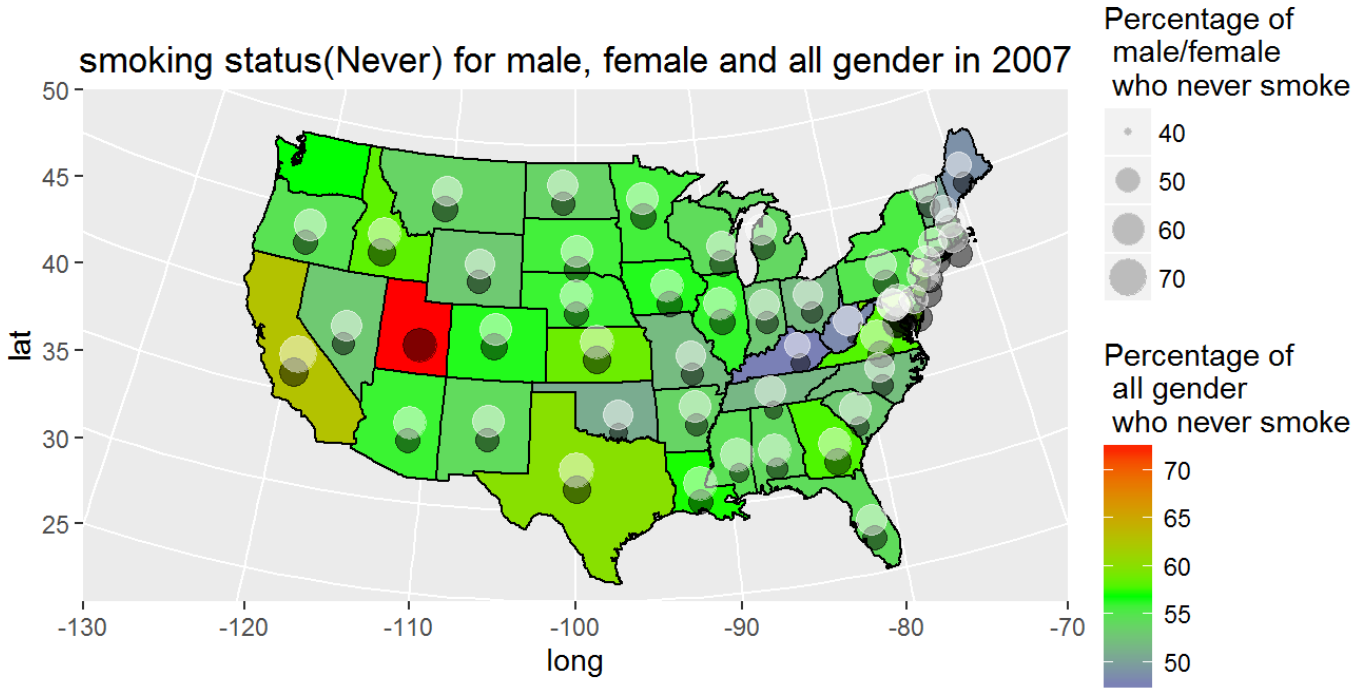



```
## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 5 rows containing missing values (geom_point).
```



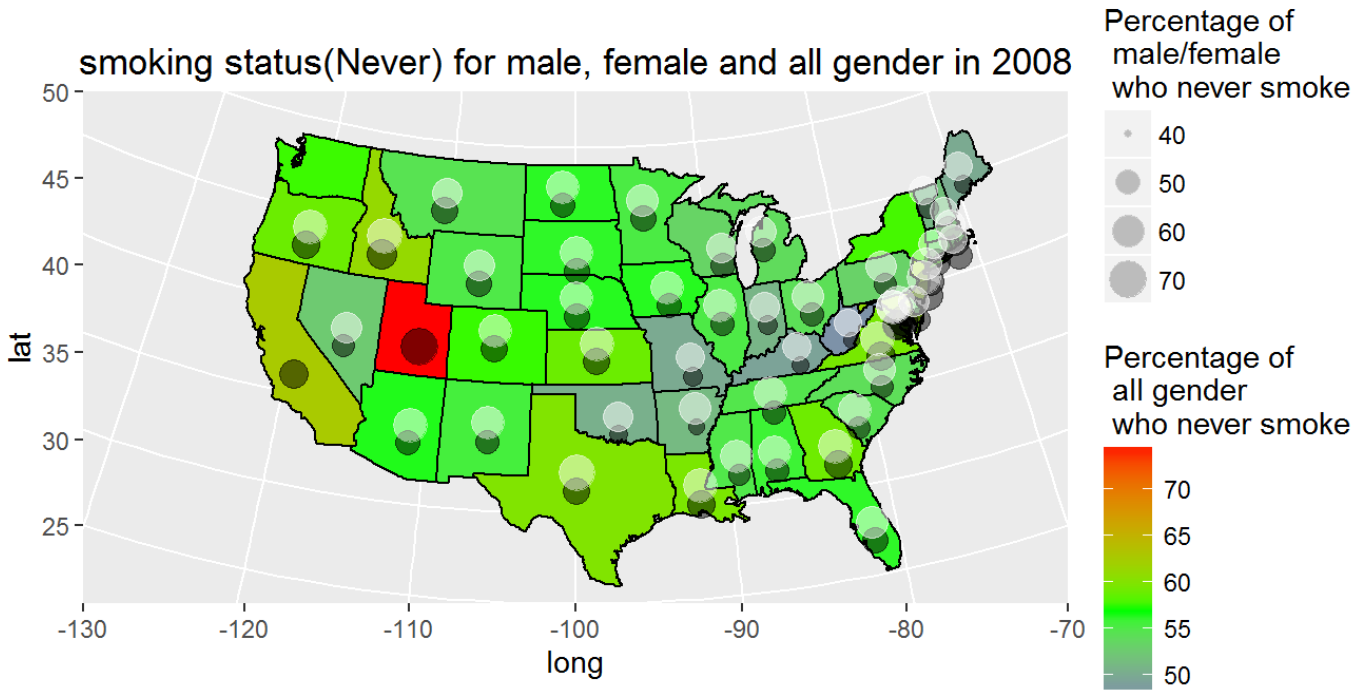

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



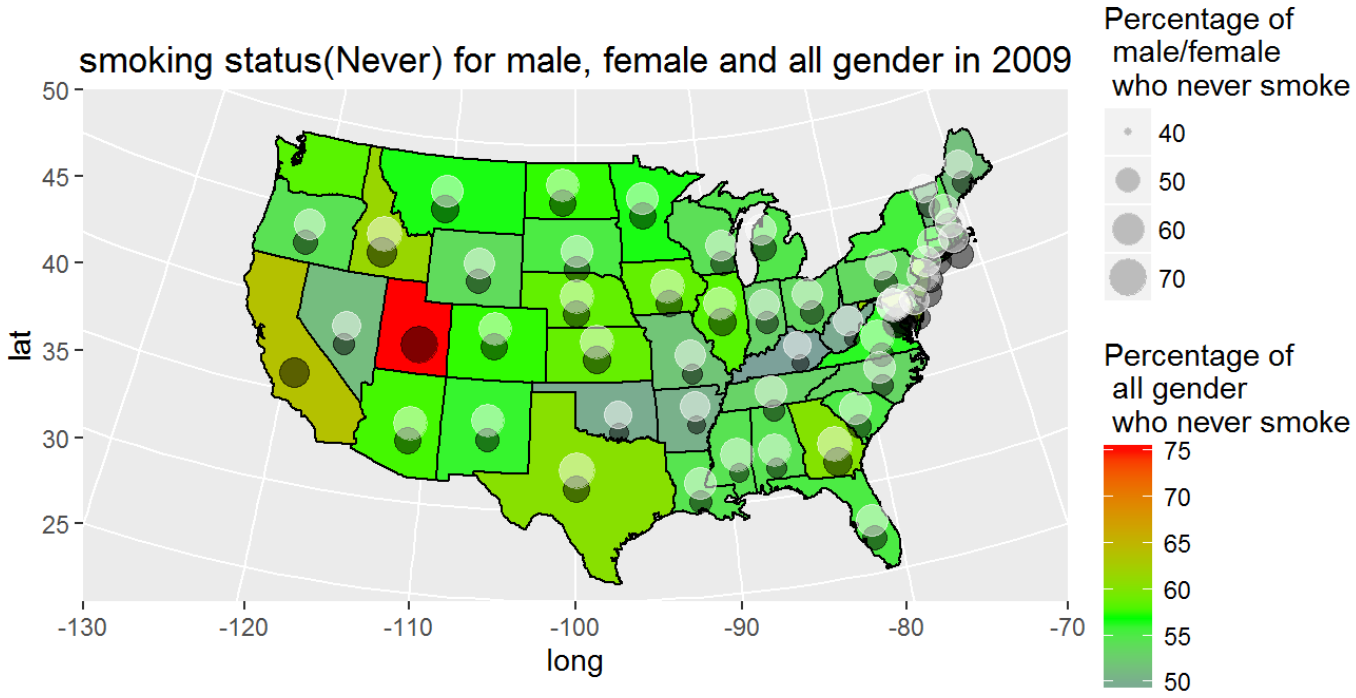

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



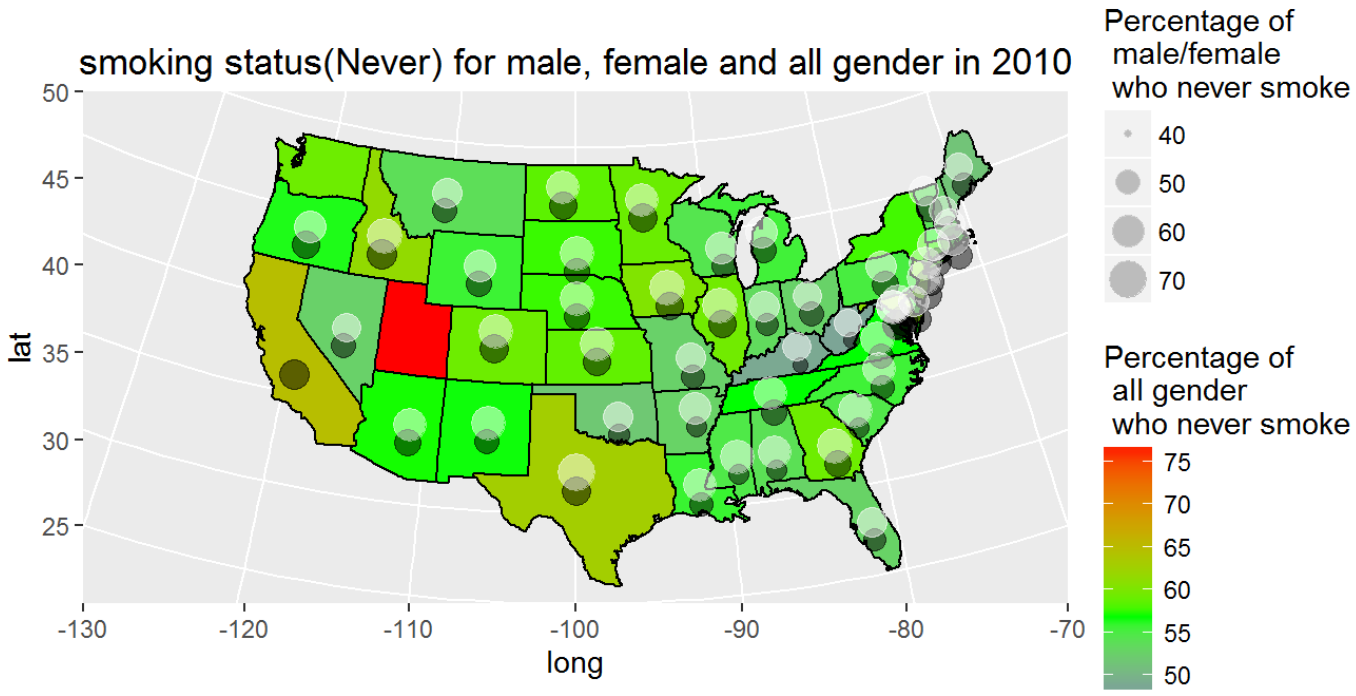

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```




```
## Warning: Removed 3 rows containing missing values (geom_point).
```

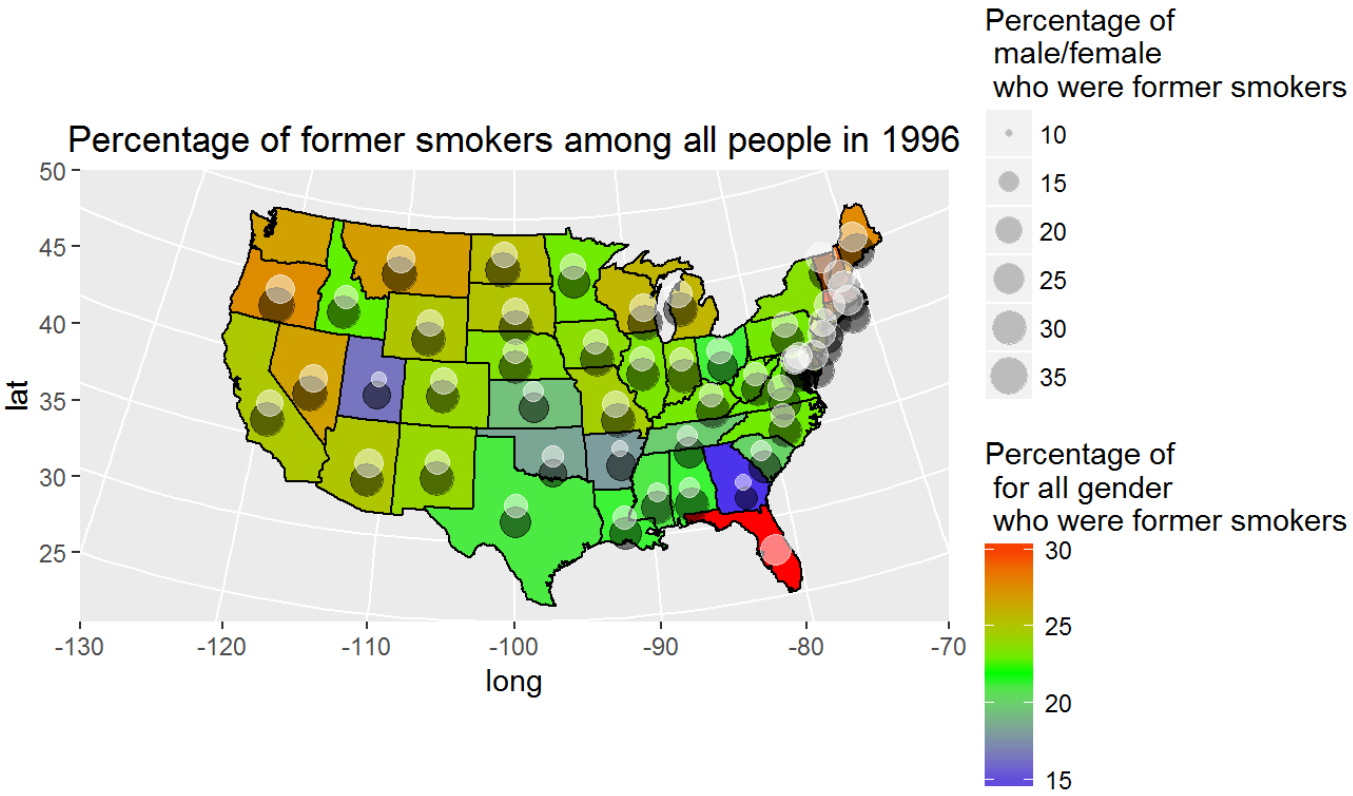
```
## Warning: Removed 5 rows containing missing values (geom_point).
```



We can tell that the percentage of people never smoke increases, so does the percentage of male never smoke. One interesting thing is that state “UT” always keeps a high percentage of people who never smoke.

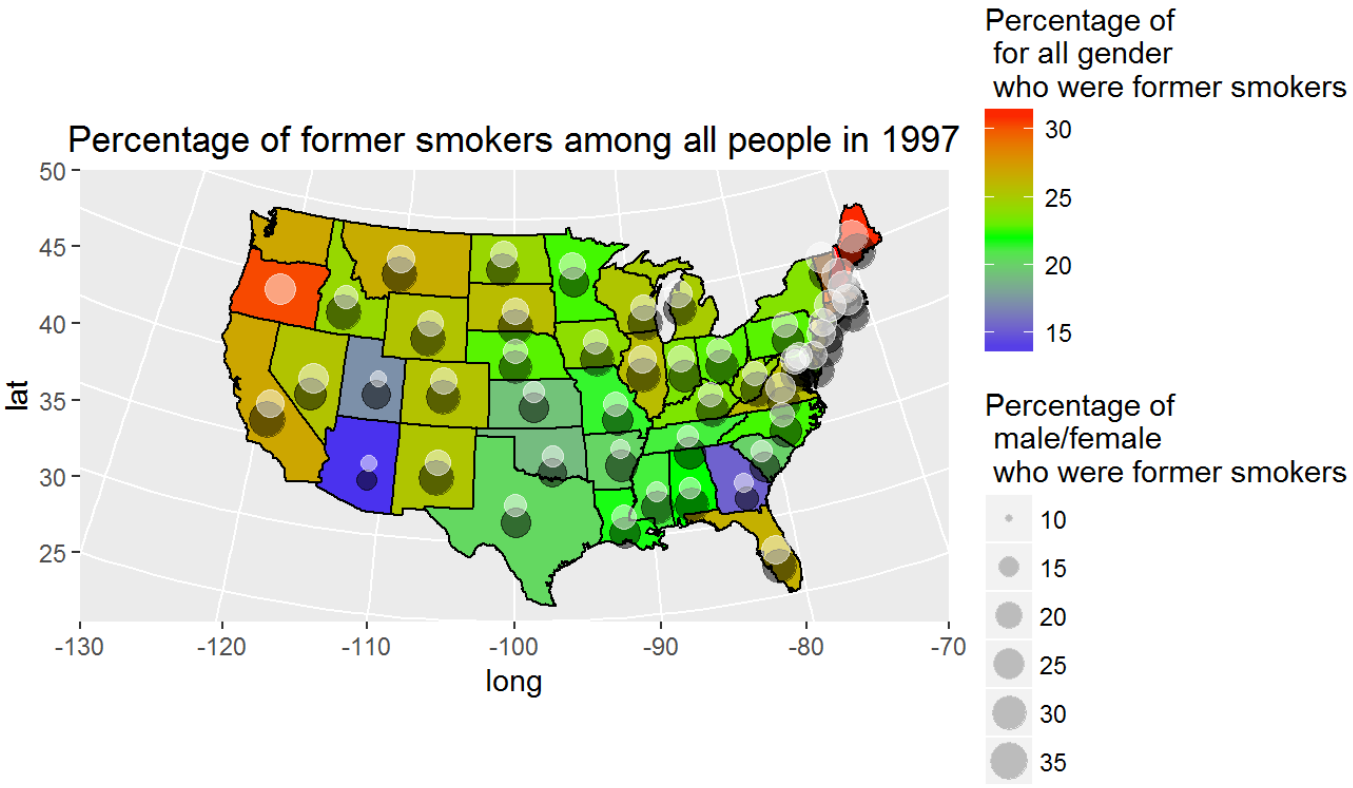
```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



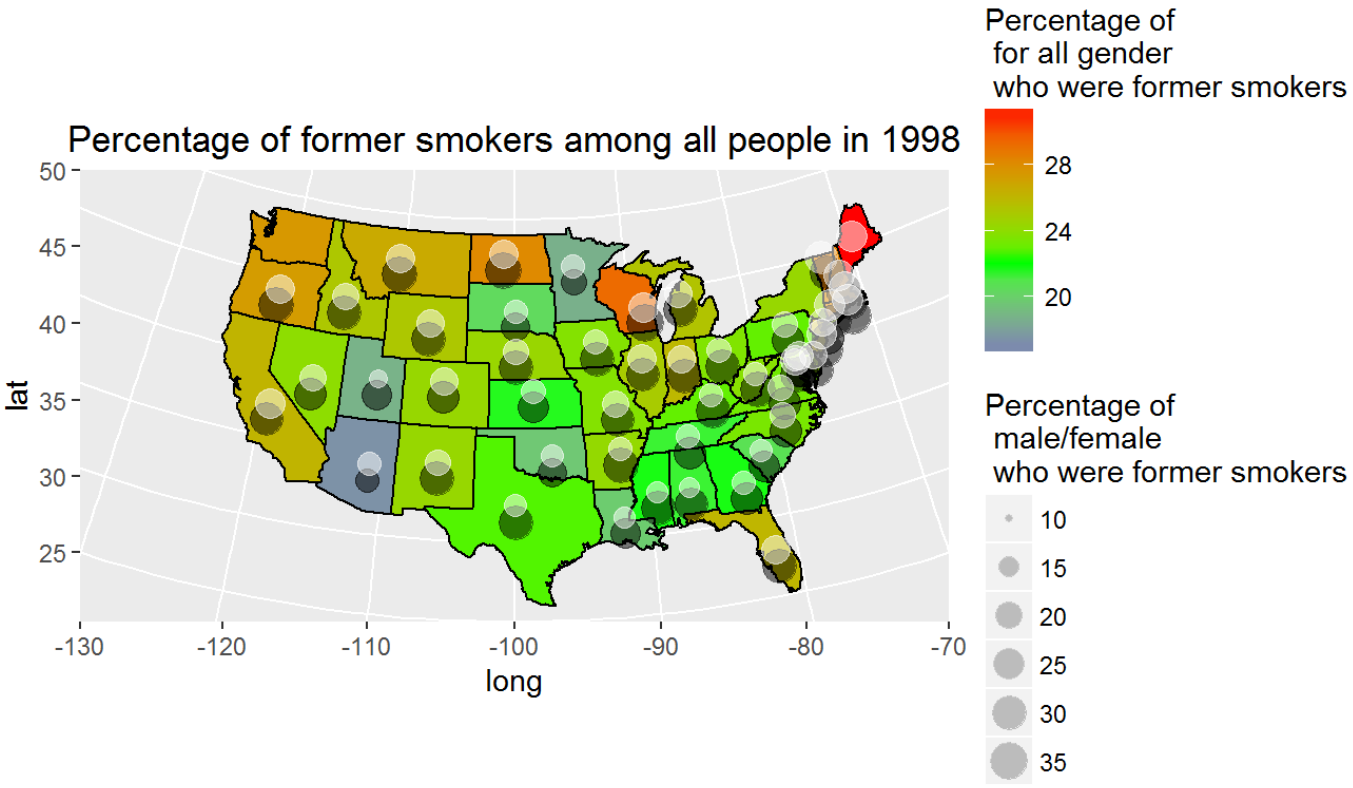

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



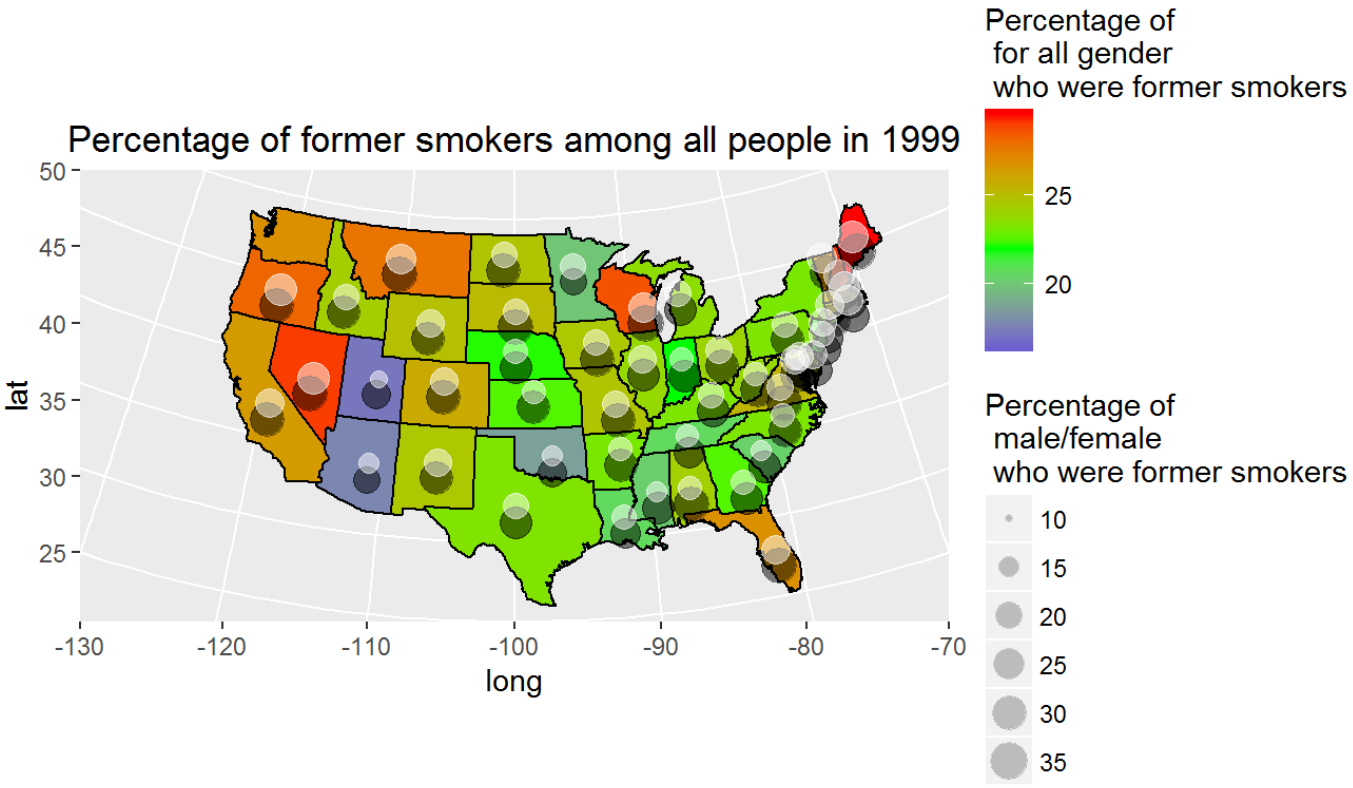

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

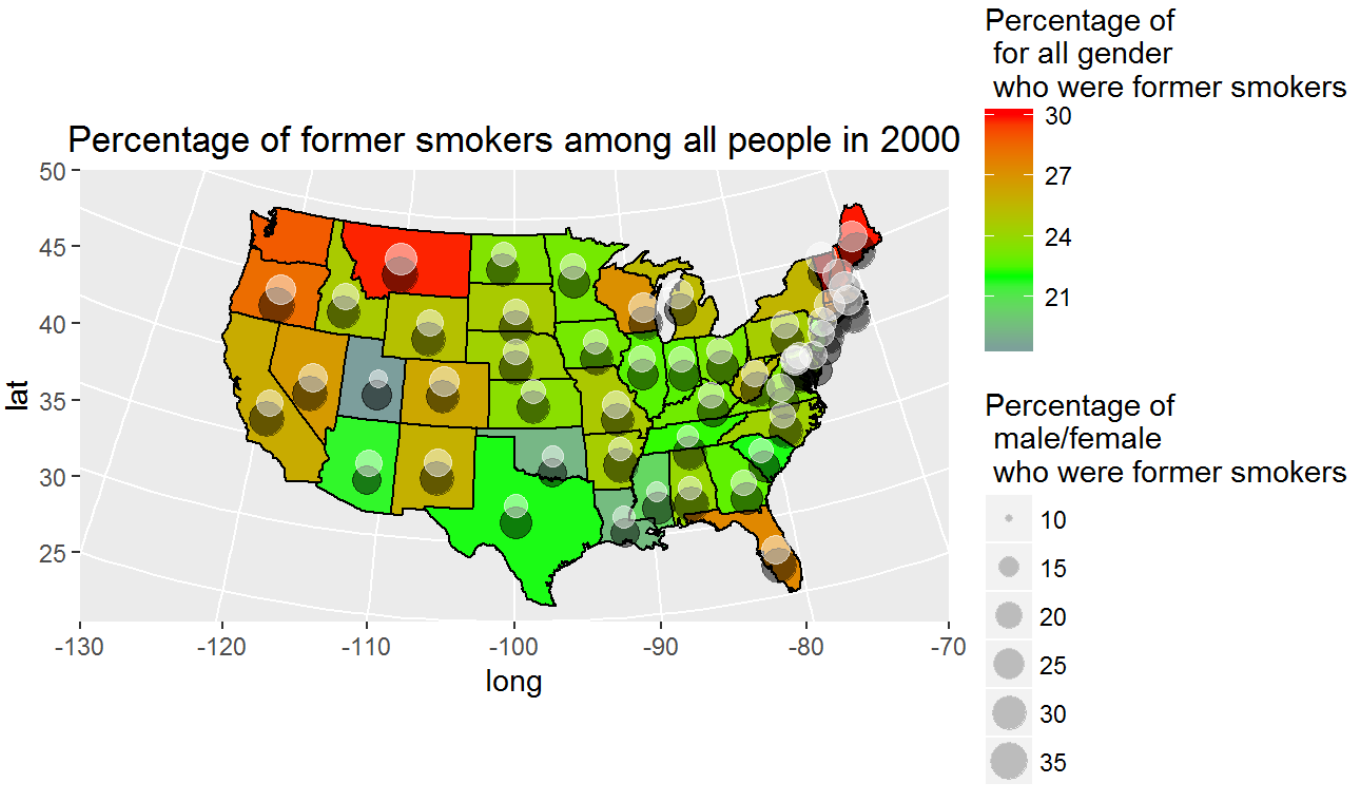



```
## Warning: Removed 3 rows containing missing values (geom_point).
```

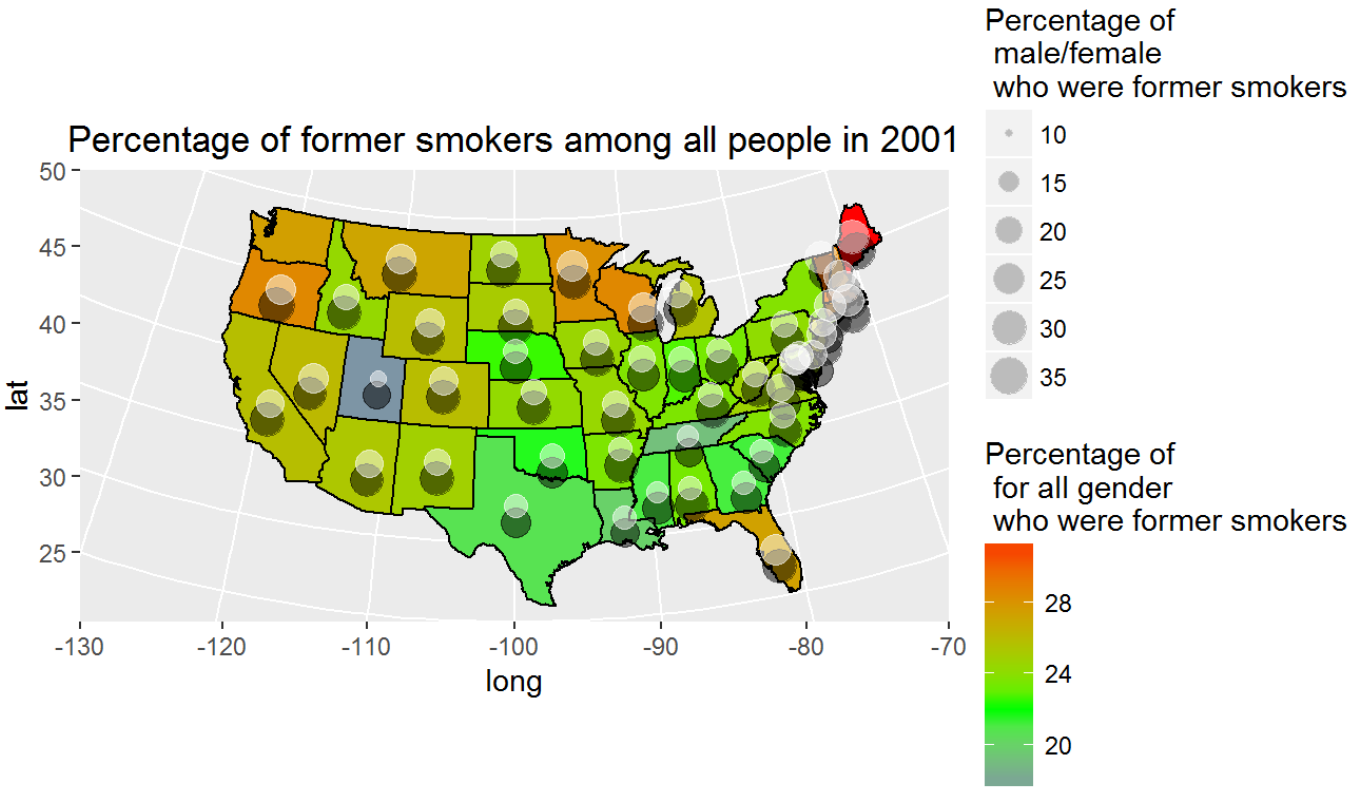
```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 4 rows containing missing values (geom_point).
```

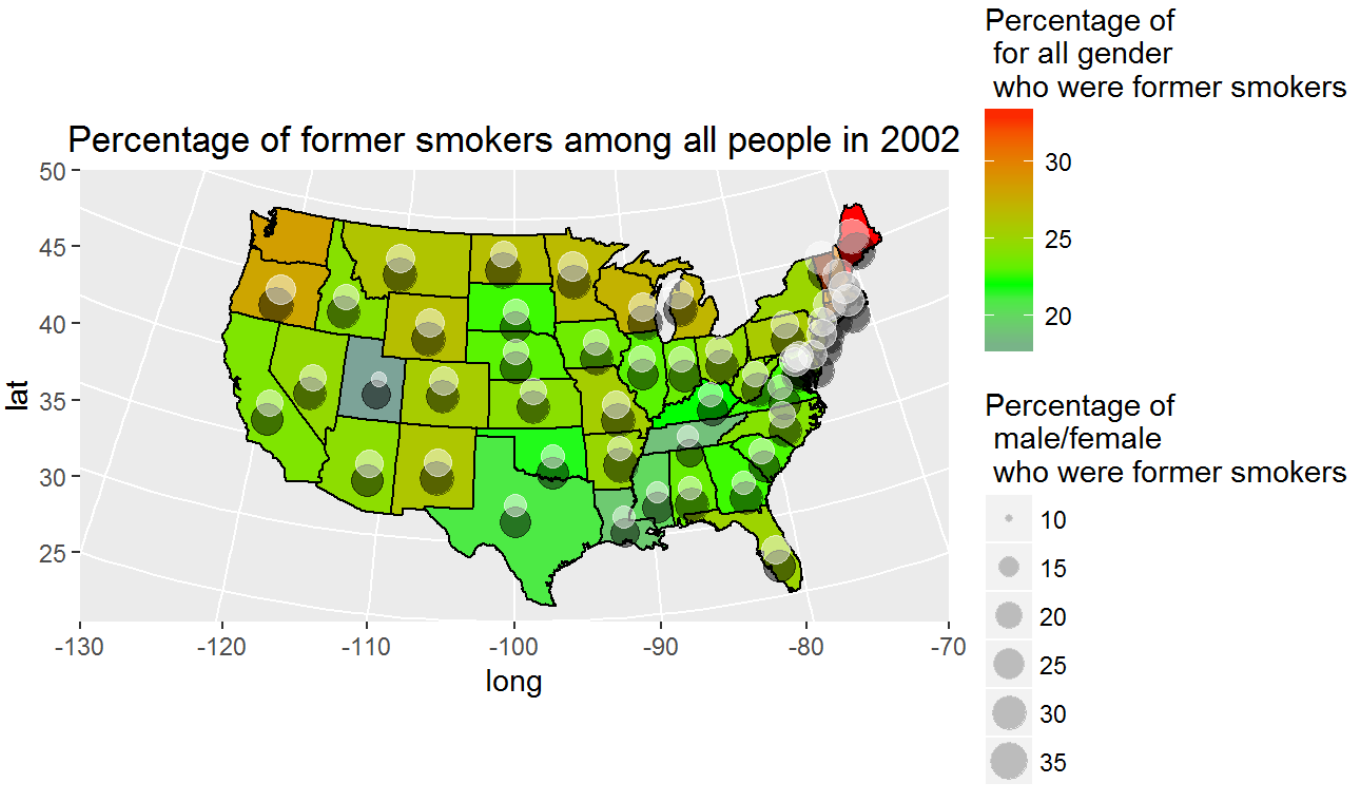



```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 4 rows containing missing values (geom_point).
```

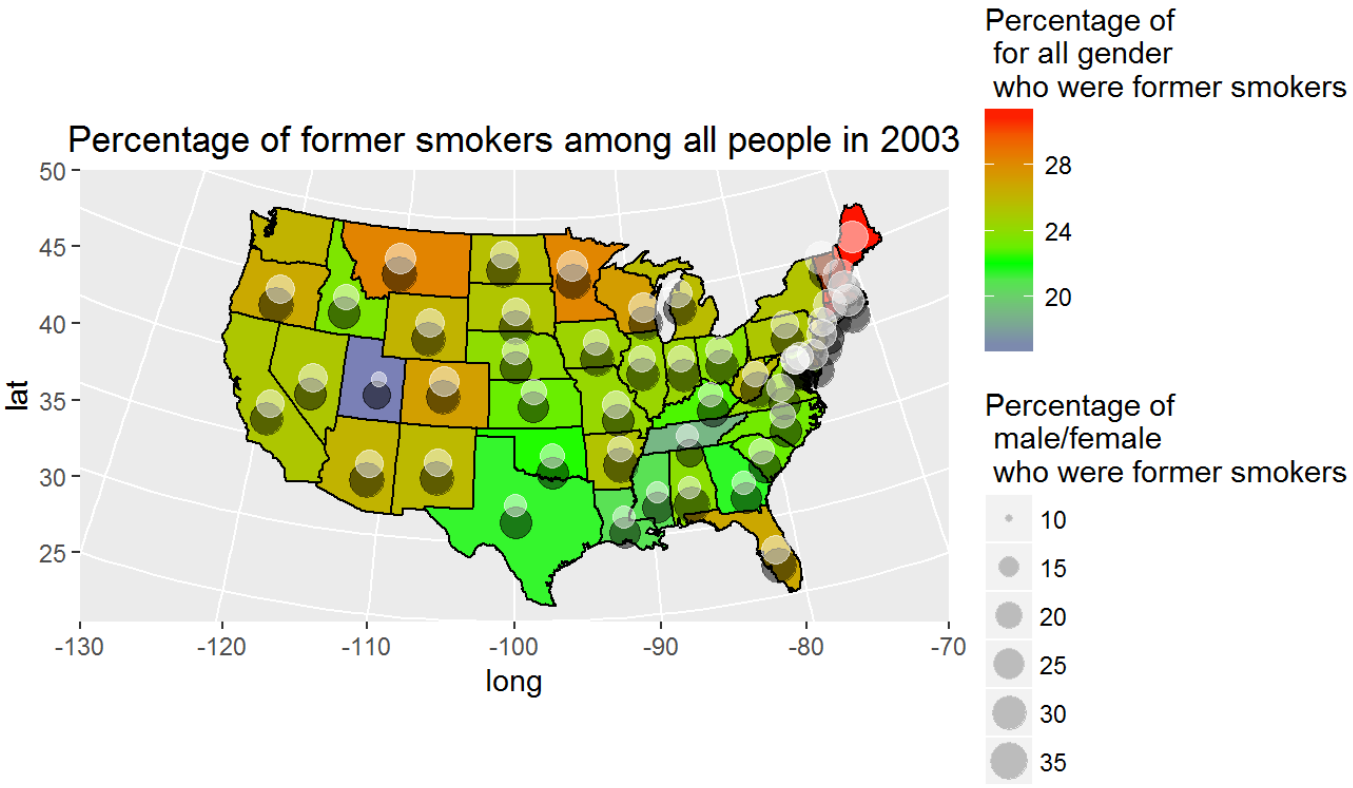



```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

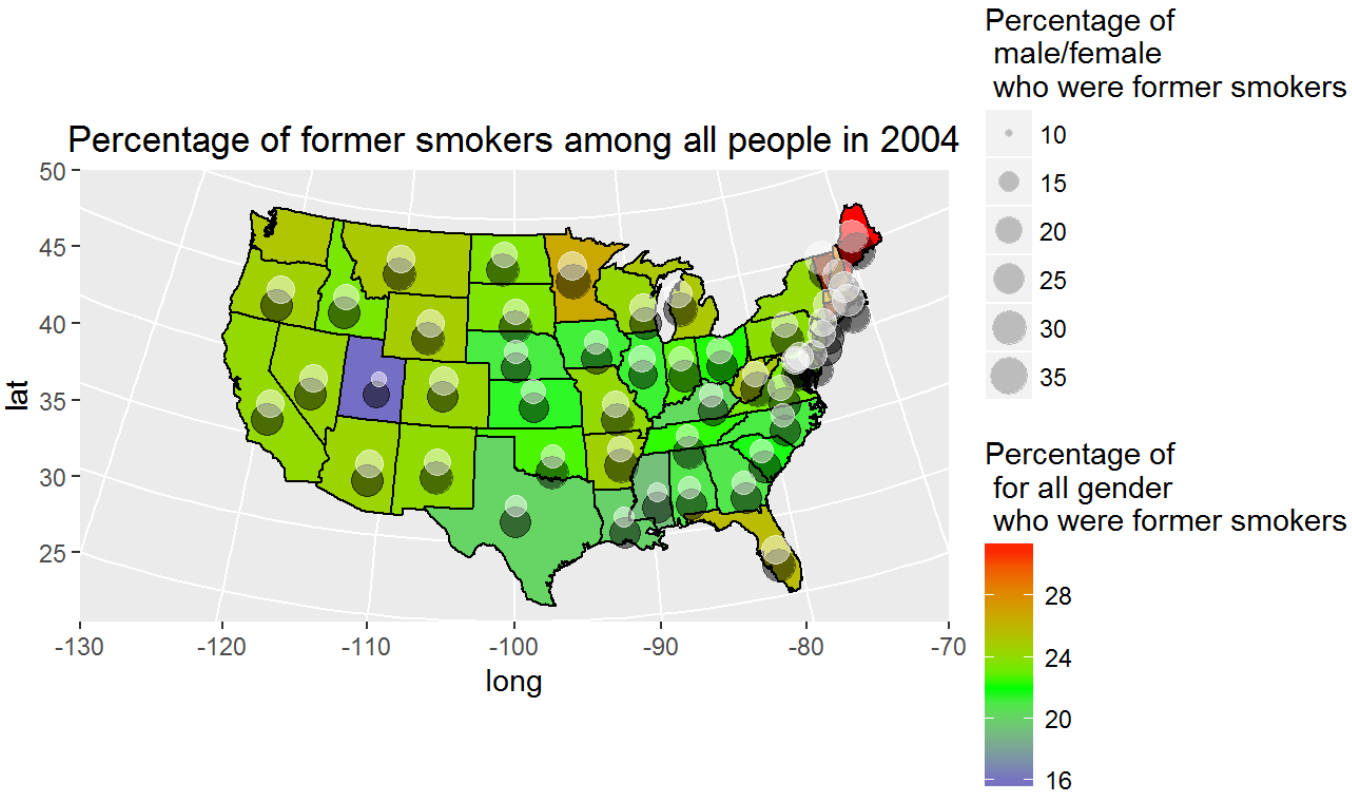



```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 2 rows containing missing values (geom_point).
```



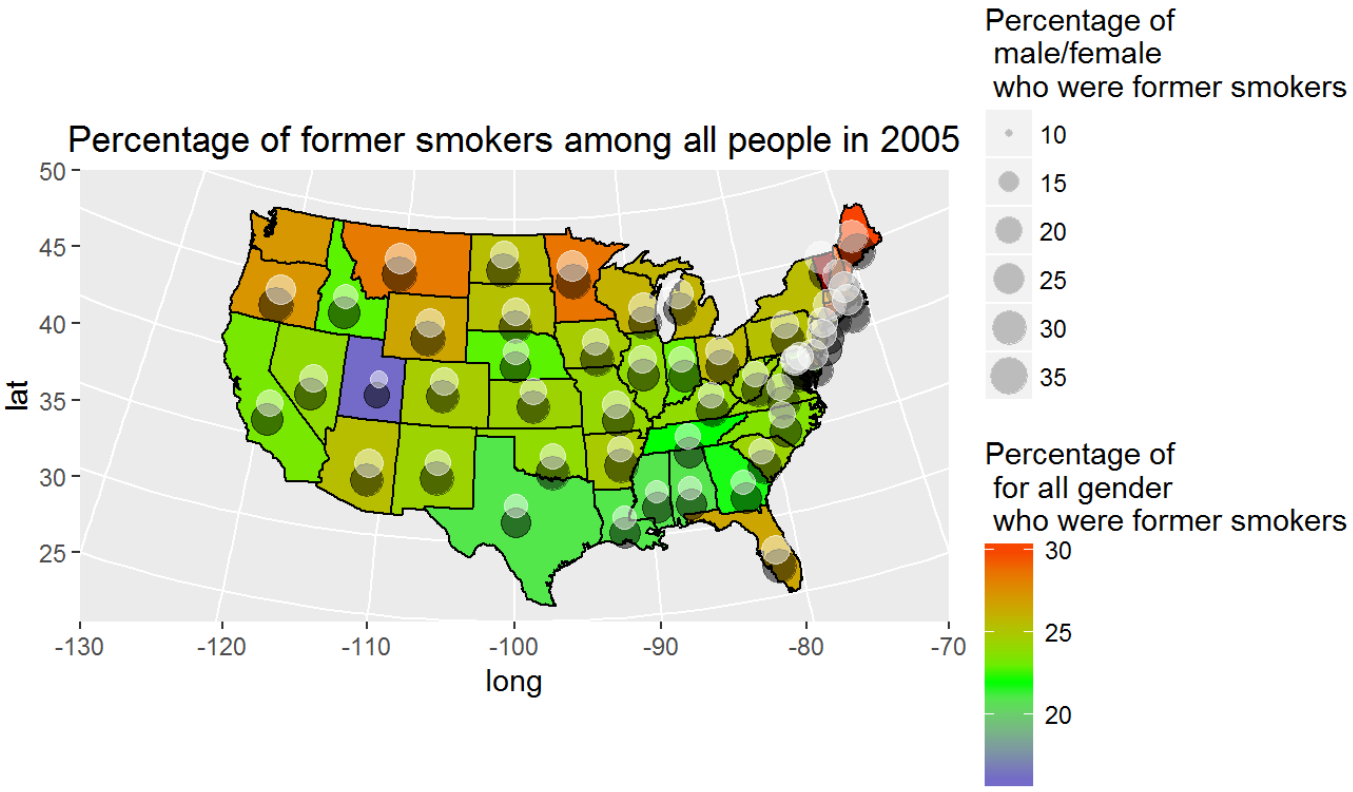

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

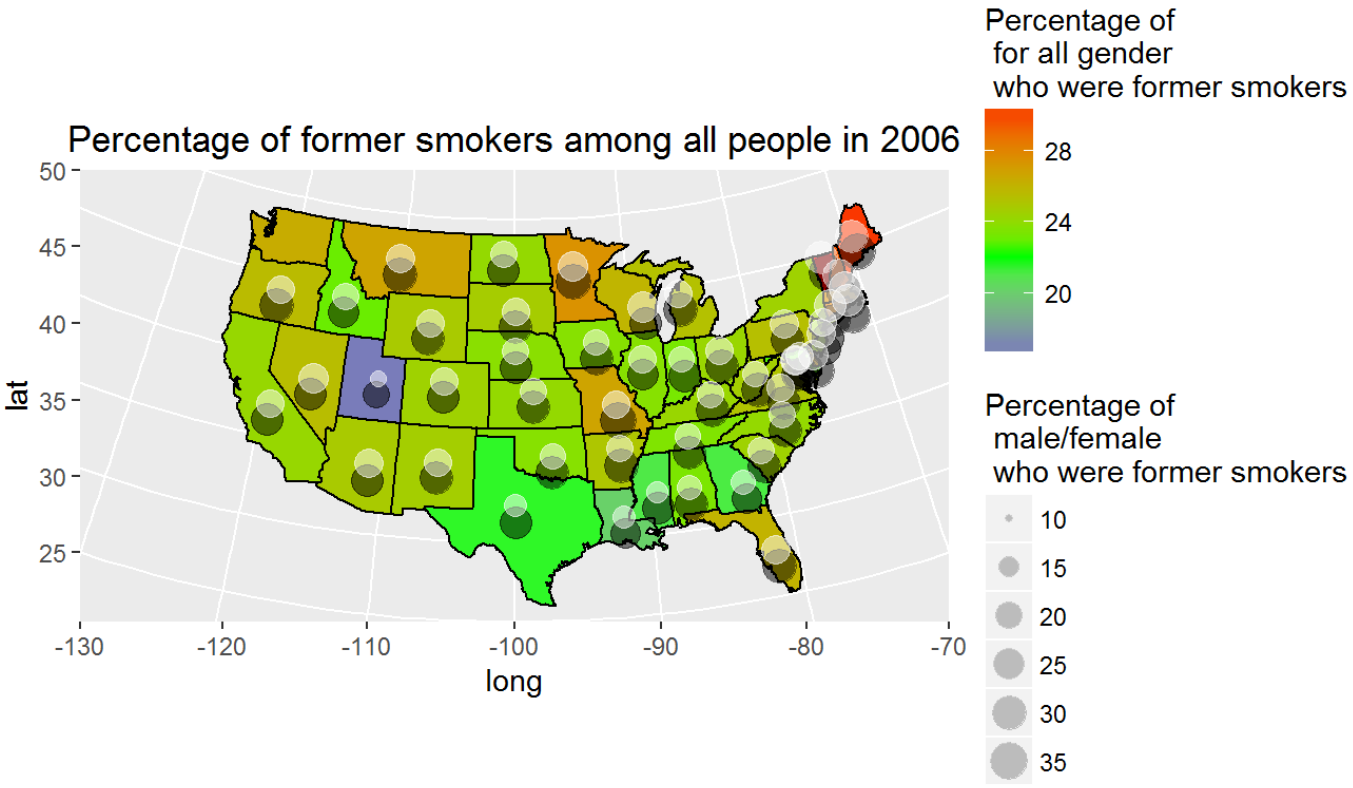



```
## Warning: Removed 2 rows containing missing values (geom_point).
```

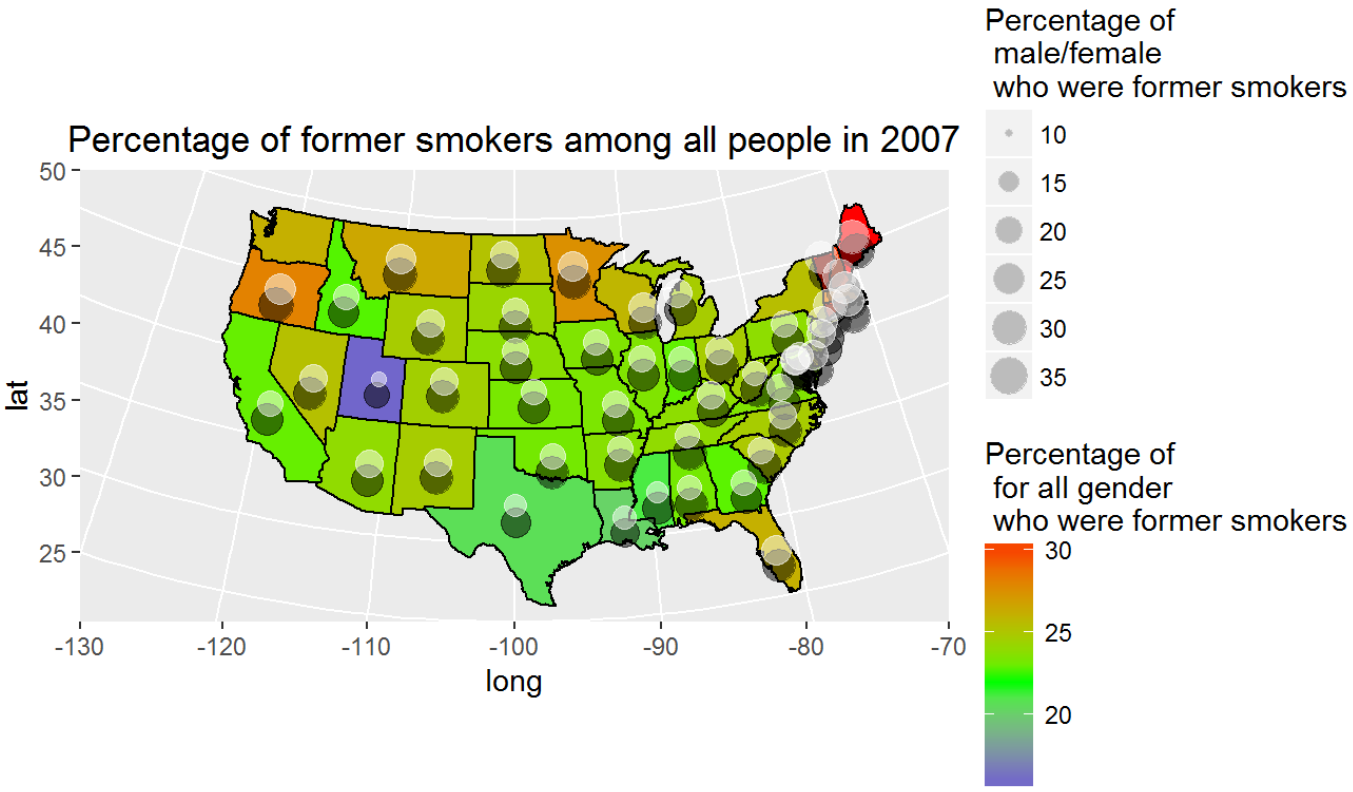
```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```
## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_point).
```

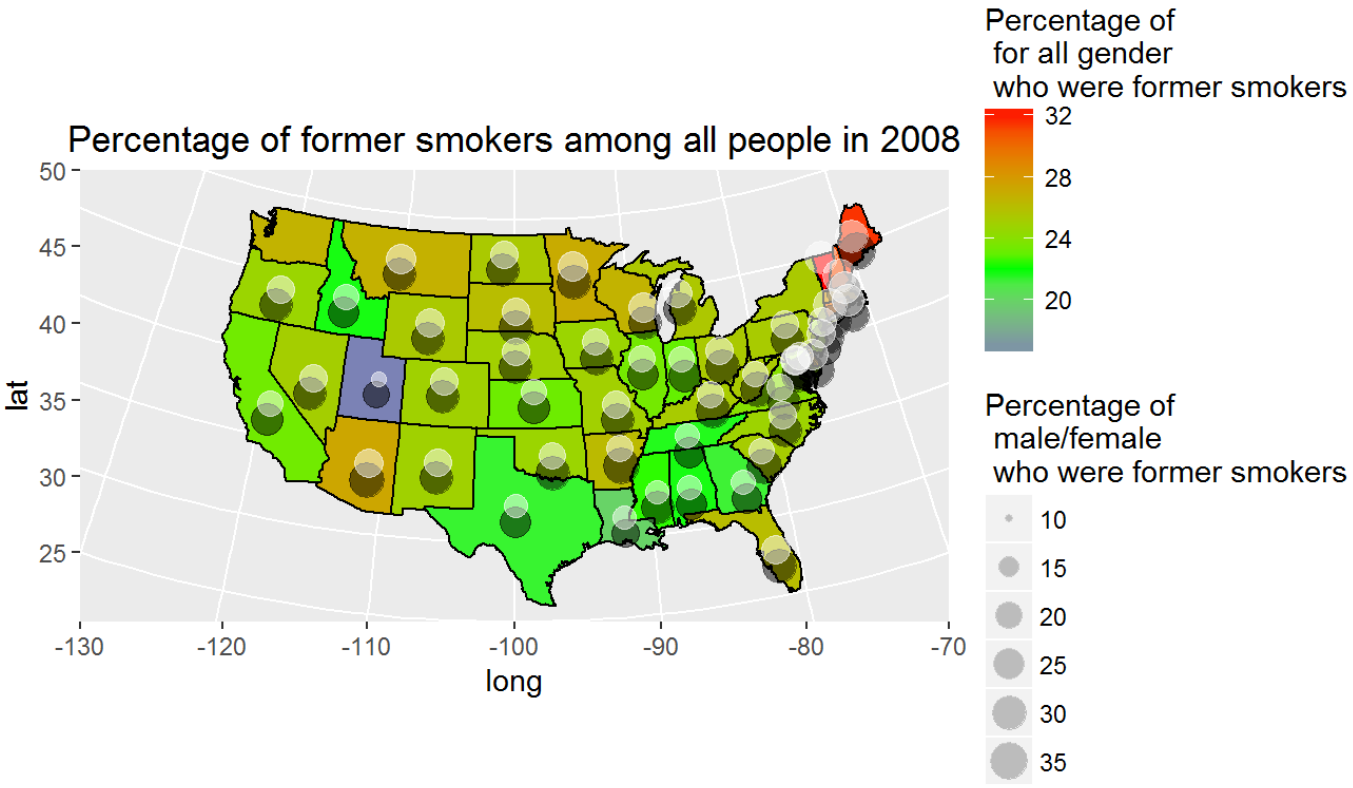



```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_point).
```



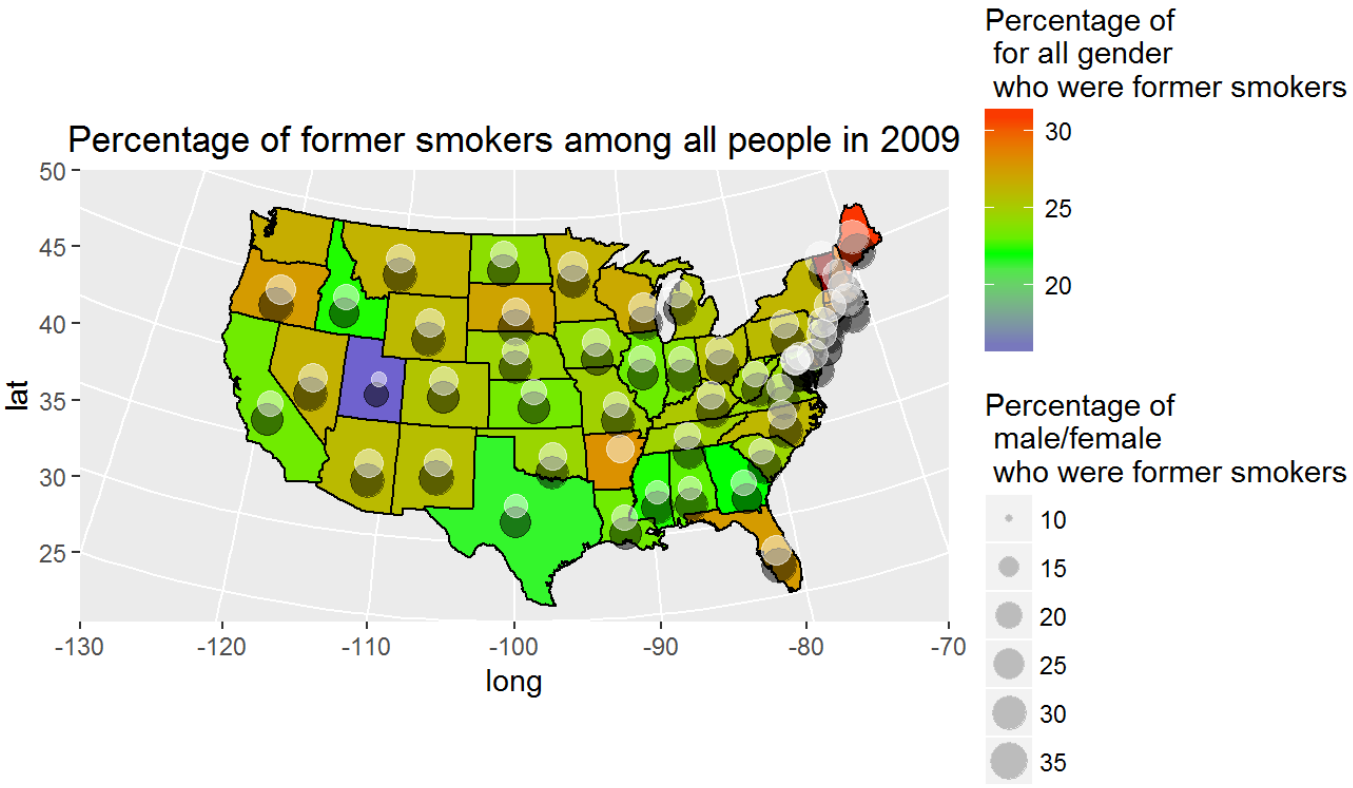

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



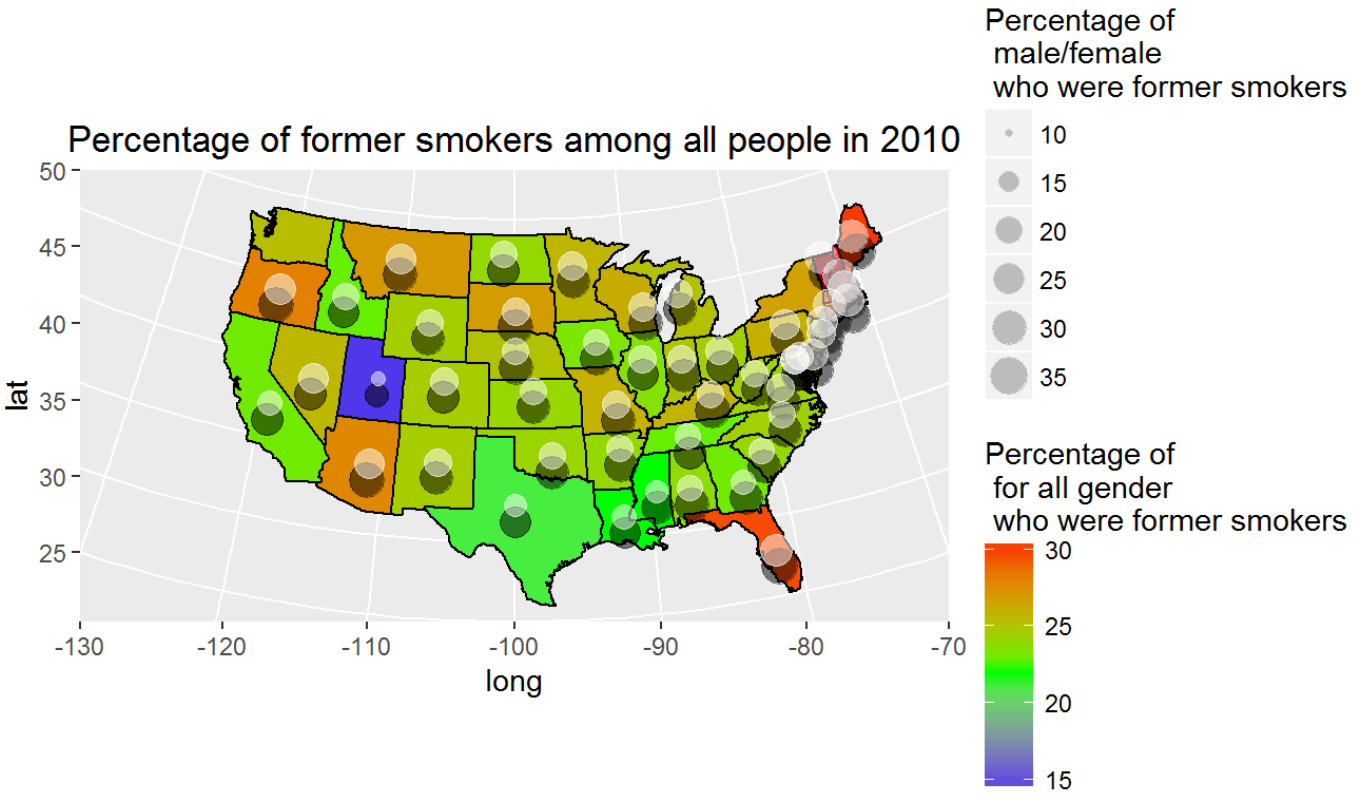

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



It is even harder to figure out an overall trend of the distribution of people are former smoker in 15 years. But we could observe that compared to 1996, states that have low percentage of people are former smokers increases in 2010. We can tell that by simply the change of color. Also, the percentage for male and female who are former smokers also increases. The percentage for male who are former smokers increases faster before 2000 and the percentage for female who are former smokers increases faster after 2002.

Discussion

The Challenge part of the project is the dataset. The dataset is different from what we have learnt before. It is recorded in the form of a controlled experiment that every time it controls one or two variables and find out how much percent of people match certain characteristics described in each row. This also limit what we can do on the dataset that we could not do validation and modelling. Therefore, we focus more on how to transform the dataset and make it clearer and how to visualize the information stored in the dataset. Another thing I think is challenging is how to visualize the dataset. The package we use to visualize it is ggmap, which we know little about it, but once we figure out how to use it, the project becomes so interesting and we have created some nice graphs. The part I think really amazing is that Tony and I figure out how to make the visualization “move”. At first, I was thinking about using R, but the “animation” package needs to download “ImageMagick” which we could not run it. SO we later thought

about Python or Java, but we both are not familiar with them. Therefore, I decided to use Imovie to make a short video, but Tony found out that there is a GIF maker website that enables us to make GIF animation. We both think that is the best way to present the visualization.

Summary

The graphs for the distribution of the Tobacco use over the US from 1996 to 2010 shows a trend that people smoke less. The visualization we made also shows the same thing. Since we all know that overdose of Tobacco would cause a harmful effect on our body, we could conclude that from the standpoint of the decrease of using in Tobacco, people are having a healthier life in 2010 than people do in 1996.

Citation

We will mark all the work used resources in the R file, following are the resources:

1. Frazier, M. (n.d.). Ggmap quickstart. Retrieved from https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf (https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf) ggmap information and usage
2. Drawing colored US State map with cut_number() in R. (n.d.). Retrieved from http://stackoverflow.com/questions/29322556/drawing-colored-us-state-map-with-cut-number-in-r (http://stackoverflow.com/questions/29322556/drawing-colored-us-state-map-with-cut-number-in-r)
3. Create a data frame of a map data. (n.d.). Retrieved from http://docs.ggplot2.org/0.9.3/map_data.html (http://docs.ggplot2.org/0.9.3/map_data.html)
4. Package ‘ggmap’. Kahle, D., & Wickham, H. (n.d.). Retrieved from https://cran.r-project.org/web/packages/ggmap/ggmap.pdf (https://cran.r-project.org/web/packages/ggmap/ggmap.pdf)
5. Save csv files. (n.d.). Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/utils/html/write.table.html (https://stat.ethz.ch/R-manual/R-devel/library/utils/html/write.table.html)