cph_738_h3

grienne

September 18, 2018

```
## -- Attaching packages -----------------------------------------------------------------------------

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ----------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

Question 1

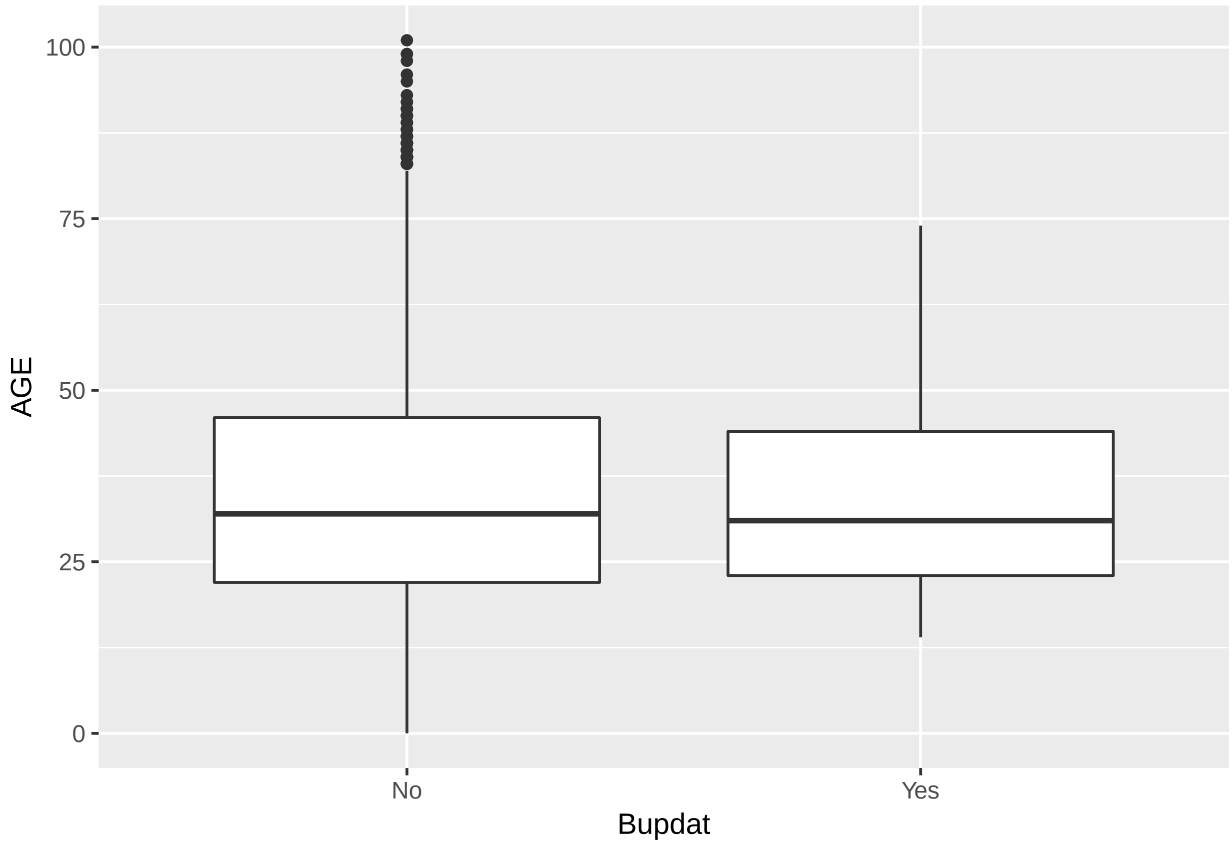
```r
##Code Bupropion 0 is No, x>0 is Yes
dat$Bupdat <- as.factor(with(dat,ifelse(Bupropion == 0, 'No', 'Yes')))

##Check to make sure still have same number of observations
summary(dat$Bupdat)
```

```
##    No   Yes
## 94909  1668
```

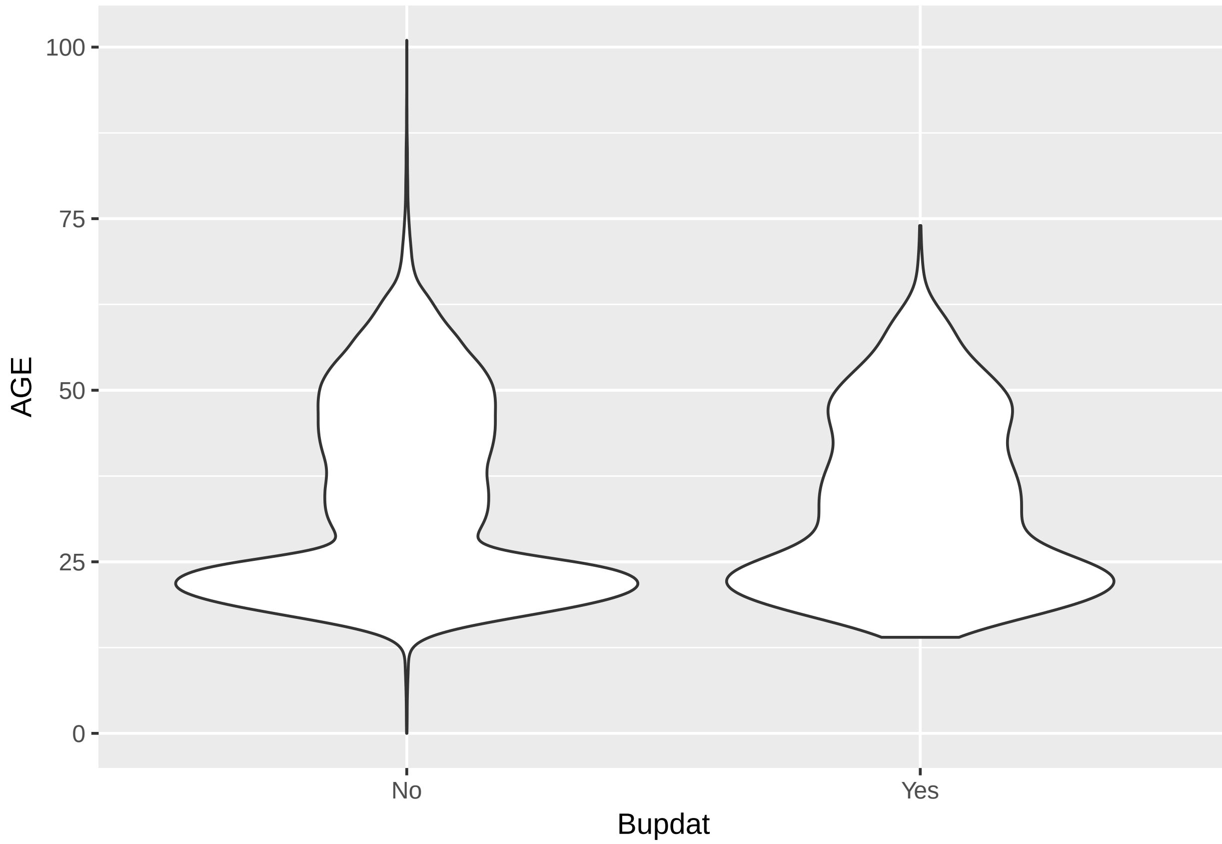
```r
str(dat$Bupdat)
```

```
##  Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

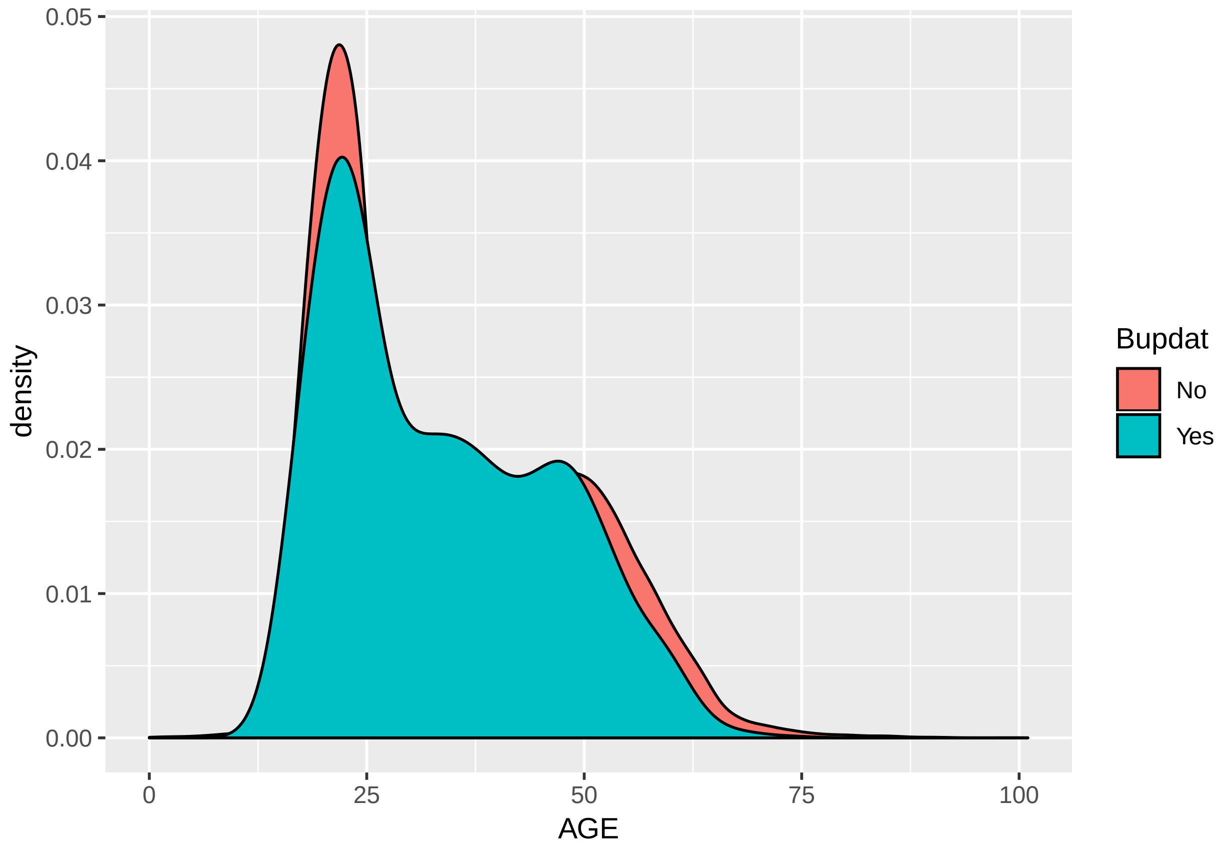
```r
##Plotting
q1 <- ggplot(data = dat, mapping = aes(y = AGE , x = Bupdat))
q1 + geom_boxplot()
```




```
q1 + geom_violin()
```




```
q1_2 <- ggplot(data = dat, mapping = aes(x = AGE , fill = Bupdat))
q1_2 + geom_density()
```



Question 2

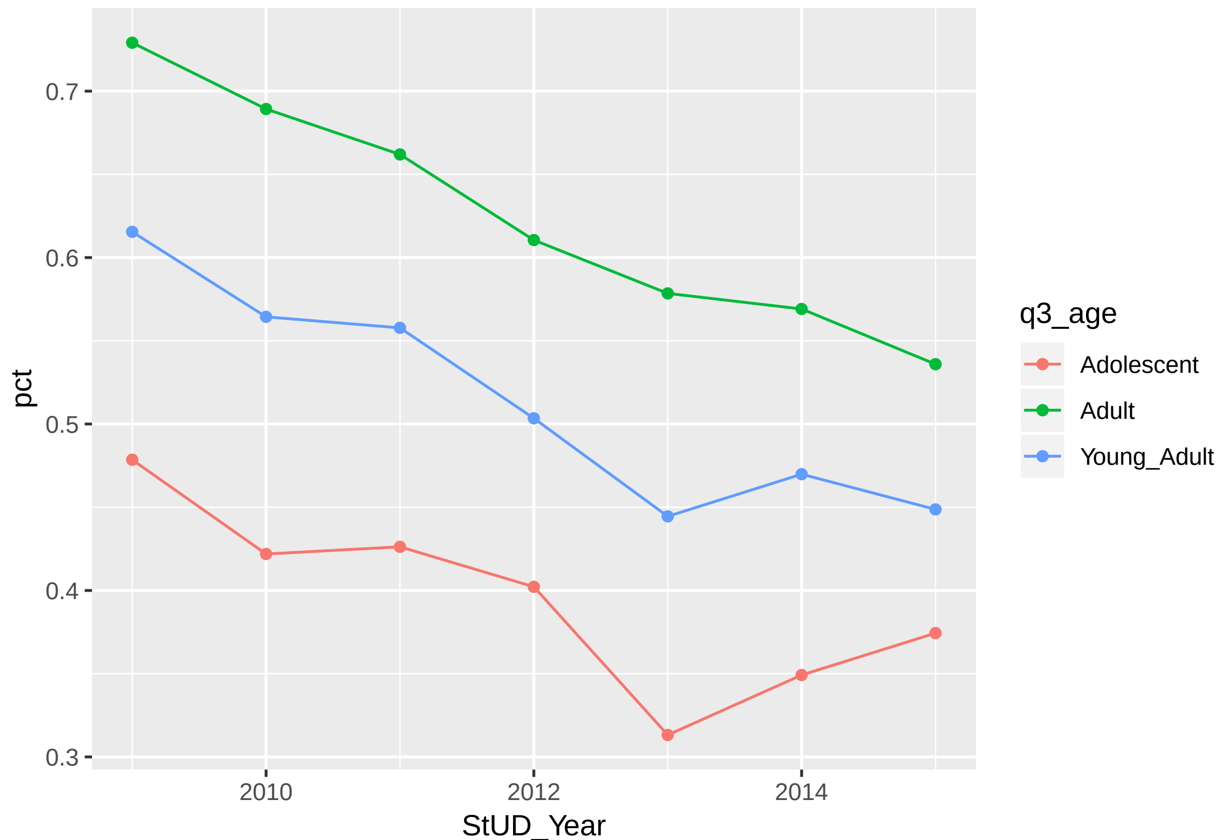
It is not appropriate to compare the average age of subjects between those who use and don't use Bupriopion. There were over 90k "no" and only 1400 "yes." Reviewing the density plot shows a bimodal distribution.

Question 3

```
#Create Age Categories
dat$q3_age <- with( dat,
                      ifelse( 12 <= AGE & AGE <= 18, "Adolescent",
                        ifelse( 19<= AGE & AGE <= 27, "Young_Adult",
                          ifelse(28 <= AGE & AGE <= 65, "Adult", "Remove")
                                )
                              )
                           )



#Create Data Set with just those in the respective Age Categories

dat$SUD_Type <- with(dat,
                      ifelse(StUD_Type == 1, "Cocaine", "All"
                                                          )
                      )
```

```r
####
q3_dat <- dat %>%
  select(StUD_Year, StUD_Type, q3_age, SUD_Type) %>%
  filter(q3_age != "Remove")


## Adolescent

q3_adol <- dat %>%
  select(q3_age,  SUD_Type, StUD_Year) %>%
  filter(q3_age == "Adolescent")

###generate percentages
q3_adol_1 <- q3_adol %>%
    group_by(q3_age, StUD_Year, SUD_Type)  %>%
      summarise(n = n()) %>%
      mutate(pct = n / sum(n))



##Young Adult

q3_yadul <- dat %>%
  select(q3_age,StUD_Year, SUD_Type) %>%
  filter(q3_age == "Young_Adult")

##Age Tables Adolescent
q3_yadul <- dat %>%
  select(q3_age,  SUD_Type, StUD_Year) %>%
  filter(q3_age == "Young_Adult")

###generate percentages
q3_yadul_1 <- q3_yadul %>%
    group_by(q3_age, StUD_Year, SUD_Type)  %>%
      summarise(n = n()) %>%
      mutate(pct = n / sum(n))


##Adult
q3_adul <- dat %>%
  select(q3_age, StUD_Year, SUD_Type) %>%
  filter(q3_age == "Adult")
##Age Tables Adolescent
q3_adul <- dat %>%
  select(q3_age,  SUD_Type, StUD_Year) %>%
  filter(q3_age == "Adult")

##generate percentages
q3_adul_1 <- q3_adul %>%
    group_by(q3_age, StUD_Year, SUD_Type)  %>%
      summarise(n = n()) %>%
      mutate(pct = n / sum(n))
```

```r
##collapse frames together
q3_final <- bind_rows(q3_adol_1, q3_adul_1, q3_yadul_1) %>%
          filter(SUD_Type != "All")


##Generate Plot

p_q3 <- ggplot(data = q3_final, mapping = aes(x = StUD_Year, y= pct, colour = q3_age))
p_q3 + geom_point() +  geom_line()
```



Question 4

The question 3 graph shows the proportion of cocaine only substance abuse disorder amongst substance use disorders. The denominator incorporates the use of amphetamines, cocaine, and those who use both amphetamines and cocaine. The graph shows that cocaine use disorder as a proportion of substance disorders has decreased since 2009 across all age groups with a slight rise in adolescents. Given the increase in prescription drug use (1) the decrease in proportional "cocaine only"" use is likely due to the changing landscape of substance use.

Most adolescents start using marijuana first (3) then prescription drug use, this conflicts with the Q3 graph which would indicate that abuse is high in cocaine, but most research focuses on the population as a whole as a result likely skewing the findings. The Truven dataset covers only about 50% of the population as a result those who are particularly disadvantaged are not represented and the demographics and behavior of initial drug use are likely to be skewed. Cocaine use disorder accounts for almost 50% of substance abuse disorders according to the graph, but the other portion covers cocaine use and amphetamine as well as amphetamine alone. The high use of cocaine is what is not consistent.

Distinguishing between the young adult and adult population can be difficult at times. However, Schulte (4) reported that young adults had the highest rates of medical emergencies due to drug use in comparison to other groups due to marijuana,a heroin, and amphetamines.SAMHSA showed that heroin was used significantly more often than cocaine (2) conflicting with the proportions found, it also showed a stabilizing rate of use which has been corroborated in other reporting (1), (3), (4). The adult population proportional cocaine use seems similar, with significantly lower cocaine use rates than seen here and a more stabilized rate overall. Overall use rates can not be determined as the graph shows proportional to overall substance use, but the downuse use trend is fairly consistent.

Overall, the trend downwards for cocaine use is similar to other reports (3), but the proportion does not seem to be accurately representative although it should be noted that the Truven set only has approximately 50% of the population so generalizability is limited. The Truven data and the question 3 graph focused on amphetamines compared to cocaine so making inferences about proportional comparison must be taken judiciously, but the overall trend downwards is accurate. Given that the graph focuses on comparing amphetamine to cocaine use the findings may balance out if other substance use disorders were incorporated into the comparison. The proportional usage does not seem consistent with national averages as cocaine is consistently significantly lower in other research findings.

1. Center for Behavioral Health Statistics and Quality. (2015). Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health (HHS Publication No. SMA 15-4927, NSDUH Series H-50). Retrieved from http://www.samhsa.gov/data/
2. Lipari, R.N. and Van Horn, S.L. Trends in substance use disorders among adults aged 18 or older. The CBHSQ Report: June 29, 2017. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.
3. NIDA. (2015, June 25). Nationwide Trends. Retrieved from https://www.drugabuse.gov/publications/drugfacts/nationwide-trends on 2018, September 22
4. Schulte, M. T., & Hser, Y.-I. (2014). Substance Use and Associated Health Conditions throughout the Lifespan. Public Health Reviews, 35(2), https://web\T1\textendashbeta.archive.org/web/20150206061220/http://www.publichealthreviews.eu/upload/pdf_files/14/00_Schulte_Hser.pdf.

Question 5

```
us_state_map <- map_data('state')
str(us_state_map)
```

```
## 'data.frame':    15537 obs. of  6 variables:
##  $ long     : num  -87.5 -87.5 -87.5 -87.5 -87.6 ...
##  $ lat      : num  30.4 30.4 30.4 30.3 30.3 ...
##  $ group    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ order    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ region   : chr  "alabama" "alabama" "alabama" "alabama" ...
##  $ subregion: chr  NA NA NA NA ...
```

```
summary(as.factor(map_data('state')$region))
```

```
##              alabama              arizona             arkansas
##                  202                  149                  312
##           california             colorado          connecticut
##                  516                   79                   91
##             delaware district of columbia              florida
##                   94                   10                  872
##              georgia                idaho             illinois
##                  381                  233                  329
```

```
##             indiana                iowa              kansas
##                 257                 256                 113
##            kentucky           louisiana               maine
##                 397                 650                 399
##            maryland       massachusetts            michigan
##                 566                 286                 830
##           minnesota         mississippi            missouri
##                 373                 382                 315
##             montana            nebraska              nevada
##                 238                 208                  70
##       new hampshire          new jersey          new mexico
##                 125                 205                  78
##            new york      north carolina        north dakota
##                 495                 782                 105
##                ohio            oklahoma              oregon
##                 238                 284                 236
##        pennsylvania        rhode island      south carolina
##                 172                  66                 304
##        south dakota           tennessee               texas
##                 166                 289                1088
##                utah             vermont            virginia
##                  59                 129                 734
##          washington       west virginia           wisconsin
##                 545                 373                 388
##             wyoming
##                  68
```

```
q5_dat_23 <- dat %>%
    select(EGEOLOC, AGE, SUD_Type) %>%
    group_by(EGEOLOC, AGE, SUD_Type)  %>%
      summarise(n = n()) %>%
        filter(AGE <= 65) %>%
          filter(AGE >= 28)

write.csv(q5_dat_23, "q5_23.csv")


##create frame with appropriate age range
q5_dat <- dat %>%
  filter(AGE <= 65) %>%
  filter(AGE >= 28)

##generate counts
q5_dat_1 <- q5_dat %>%
  select(EGEOLOC, SUD_Type) %>%
    group_by(EGEOLOC, SUD_Type)  %>%
      summarise(n = n()) %>%
        set_names("EGEOLOC", "drug", "n")

##name states for manipulation
q5_dat_1$region <- with(q5_dat_1,
                  ifelse(EGEOLOC == 41, 'alabama',
                  ifelse(EGEOLOC == 52, 'arizona',
                  ifelse(EGEOLOC == 46, 'arkansas',
```

```
                   ifelse(EGEOLOC == 62, 'california',
                   ifelse(EGEOLOC == 53, 'colorado',
                   ifelse(EGEOLOC == 4,  'connecticut',
                   ifelse(EGEOLOC == 32, 'delaware',
                   ifelse(EGEOLOC == 31, 'district of columbia',
                   ifelse(EGEOLOC == 33, 'florida',
                   ifelse(EGEOLOC == 34, 'georgia',
                   ifelse(EGEOLOC == 54, 'idaho',
                   ifelse(EGEOLOC == 16, 'illinois',
                   ifelse(EGEOLOC == 17, 'indiana',
                   ifelse(EGEOLOC == 22, 'iowa',
                   ifelse(EGEOLOC == 23, 'kansas',
                   ifelse(EGEOLOC == 42, 'kentucky',
                   ifelse(EGEOLOC == 47, 'lousiana',
                   ifelse(EGEOLOC ==  5, 'maine',
                   ifelse(EGEOLOC == 35, 'maryland',
                   ifelse(EGEOLOC ==  6, 'massachusetts',
                   ifelse(EGEOLOC == 18, 'michigan',
                   ifelse(EGEOLOC == 24, 'minnesota',
                   ifelse(EGEOLOC == 43, 'mississippi',
                   ifelse(EGEOLOC == 25, 'missouri',
                   ifelse(EGEOLOC == 55, 'montana',
                   ifelse(EGEOLOC == 26, 'nebraska',
                   ifelse(EGEOLOC == 56, 'nevade',
                   ifelse(EGEOLOC ==  7, 'new hampshire',
                   ifelse(EGEOLOC == 11, 'new jersey',
                   ifelse(EGEOLOC == 57, 'new mexico',
                   ifelse(EGEOLOC == 12, 'new york',
                   ifelse(EGEOLOC == 36, 'north carolina',
                   ifelse(EGEOLOC == 27, 'north dakota',
                   ifelse(EGEOLOC == 19, 'ohio',
                   ifelse(EGEOLOC == 48, 'oklahoma',
                   ifelse(EGEOLOC == 64, 'oregon',
                   ifelse(EGEOLOC == 13, 'pennsylvania',
                   ifelse(EGEOLOC ==  8, 'rhode island',
                   ifelse(EGEOLOC == 37, 'south carolina',
                   ifelse(EGEOLOC == 28, 'south dakota',
                   ifelse(EGEOLOC == 44, 'tennessee',
                   ifelse(EGEOLOC == 49, 'texas',
                   ifelse(EGEOLOC == 58, 'utah',
                   ifelse(EGEOLOC ==  9, 'vermont',
                   ifelse(EGEOLOC == 38, 'virginia',
                   ifelse(EGEOLOC == 65, 'washington',
                   ifelse(EGEOLOC == 39, 'west virginia',
                   ifelse(EGEOLOC == 20, 'wisconsin',
                   ifelse(EGEOLOC == 59, 'wyoming',
                          NA))))))))))))))))))))))))))))))))))))))))))))))))))

##find proportions
q5_dat_1_final <- q5_dat_1 %>%
  select(region, drug, n) %>%
    mutate(prop = n / sum(n))
```

```
## Adding missing grouping variables: `EGEOLOC`
```

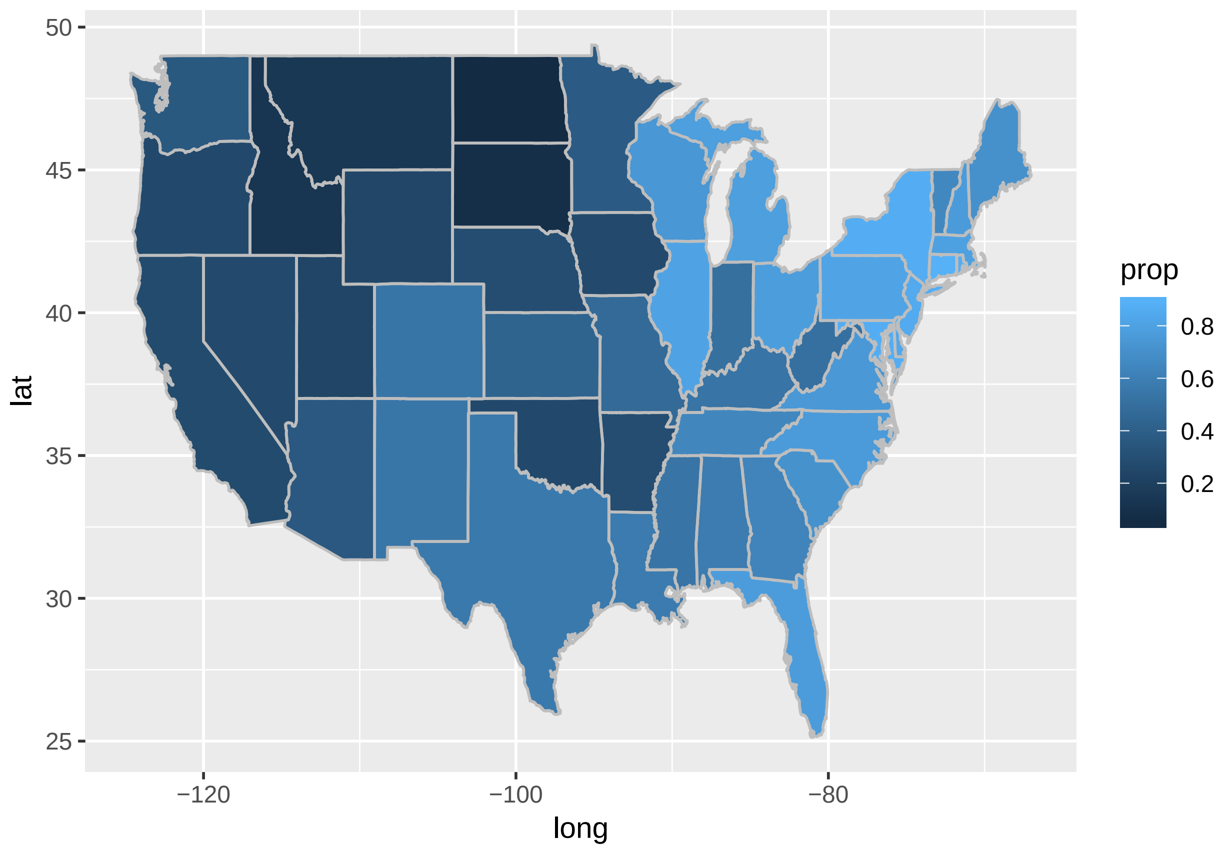
```r
##collapse states
aggregate(n ~ region, data = q5_dat_1_final, FUN = sum)
```

```
##                  region    n
## 1               alabama  539
## 2               arizona  694
## 3              arkansas  234
## 4            california 6416
## 5              colorado  471
## 6           connecticut 1333
## 7              delaware  213
## 8  district of columbia   63
## 9               florida 2223
## 10              georgia 1873
## 11                idaho  251
## 12             illinois 2286
## 13              indiana 1713
## 14                 iowa  358
## 15               kansas  245
## 16             kentucky  778
## 17             lousiana 1273
## 18                maine  162
## 19             maryland  550
## 20        massachusetts  795
## 21             michigan 1593
## 22            minnesota  396
## 23          mississippi  331
## 24             missouri  771
## 25              montana  113
## 26             nebraska  120
## 27               nevade  345
## 28        new hampshire  178
## 29           new jersey  866
## 30           new mexico  753
## 31             new york 6324
## 32       north carolina 1329
## 33         north dakota   38
## 34                 ohio 2437
## 35             oklahoma  601
## 36               oregon  436
## 37         pennsylvania 2999
## 38         rhode island  143
## 39       south carolina 1981
## 40         south dakota   36
## 41            tennessee  888
## 42                texas 4444
## 43                 utah  290
## 44              vermont   40
## 45             virginia  940
## 46           washington  817
## 47        west virginia  368
## 48            wisconsin  625
## 49              wyoming   63
```

```r
##success, note EGEOLOC will add due to it being a "grouped factor above"
##Next stage creates a temp data set that undoes the impact of "grouping that was likely causing errors
##In the future use function 'ungroup()'
q5_dat_final <- q5_dat_1_final %>%
      select(region, drug, prop) %>%
        filter(drug != "All") %>%
        filter(region != "NA")
```

```
## Adding missing grouping variables: `EGEOLOC`
```

```r
##Had to arrange by region so when pulling prop would be in correct order
##In the future simply ungrouping in the first place would have removed most of these extra steps
q5_final <- arrange(q5_dat_final, region)




## list of states
states = c("alabama","arizona","arkansas","california",
           "colorado","connecticut","delaware","district of columbia",
           "florida","georgia","idaho","illinois",
           "indiana","iowa","kansas","kentucky",
           "louisiana","maine","maryland","massachusetts",
           "michigan","minnesota","mississippi","missouri",
           "montana","nebraska","nevada","new hampshire",
           "new jersey","new mexico","new york","north carolina",
           "north dakota","ohio","oklahoma","oregon",
           "pennsylvania","rhode island","south carolina","south dakota",
           "tennessee","texas","utah","vermont",
           "virginia","washington","west virginia","wisconsin",
           "wyoming")

##creating map-useful dataset
##Meet with Olga to discuss map commands in-depth

##This creates a proportion only dataframe
prop2 <- data.frame(q5_final$prop)

##this creates a new dataframe with prop and states, changed to region
##likely this has to be done to undo the EGEOLOC grouping.
tmp_dat_q5 <- data.frame(states, prop2)
names(tmp_dat_q5) <- c('region', 'prop')

## merge with state map data
map_dat_prop_tp <- merge(us_state_map, tmp_dat_q5, by ='region', all = T)

## plot
p <- ggplot(map_dat_prop_tp, aes(x = long, y = lat, group = group))
p + geom_polygon(aes(fill = prop)) +
  geom_path(colour = 'gray')
```



Question 6

States with the highest CUD Prevalence - Conneticut, Rhode Island, Maryland States with the lowest CUD Prevalence - North Dakota, South Dakota, Idaho