MATH 208 Assignment2

Yan Miao

2019-10-07

Question 1

Load libraries and data:

```r
library(fivethirtyeight)
library(ggplot2)
library(gridExtra)
library(ggmosaic)
library(tidyverse)
data(biopics)
```

Data exploration:

```r
summary(biopics)
```

```
##     title               site              country           year_release 
##  Length:761         Length:761         Length:761         Min.   :1915  
##  Class :character   Class :character   Class :character   1st Qu.:1969  
##  Mode  :character   Mode  :character   Mode  :character   Median :1995  
##                                                           Mean   :1987  
##                                                           3rd Qu.:2007  
##                                                           Max.   :2014  
##                                                                         
##    box_office          director         number_of_subjects
##  Min.   :     3150   Length:761         Min.   :1.000     
##  1st Qu.:  1170000   Class :character   1st Qu.:1.000     
##  Median :  6140000   Mode  :character   Median :1.000     
##  Mean   : 22981174                      Mean   :1.268     
##  3rd Qu.: 30500000                      3rd Qu.:1.000     
##  Max.   :350000000                      Max.   :4.000     
##  NA's   :324                                              
##    subject          type_of_subject     race_known       
##  Length:761         Length:761         Length:761        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##                                                          
##  subject_race       person_of_color subject_sex        lead_actor_actress
##  Length:761         Mode :logical   Length:761         Length:761        
##  Class :character   FALSE:661       Class :character   Class :character  
##  Mode  :character   TRUE :100       Mode  :character   Mode  :character  
##                                                                          
##                                                                          
##                                                                          
## 
```

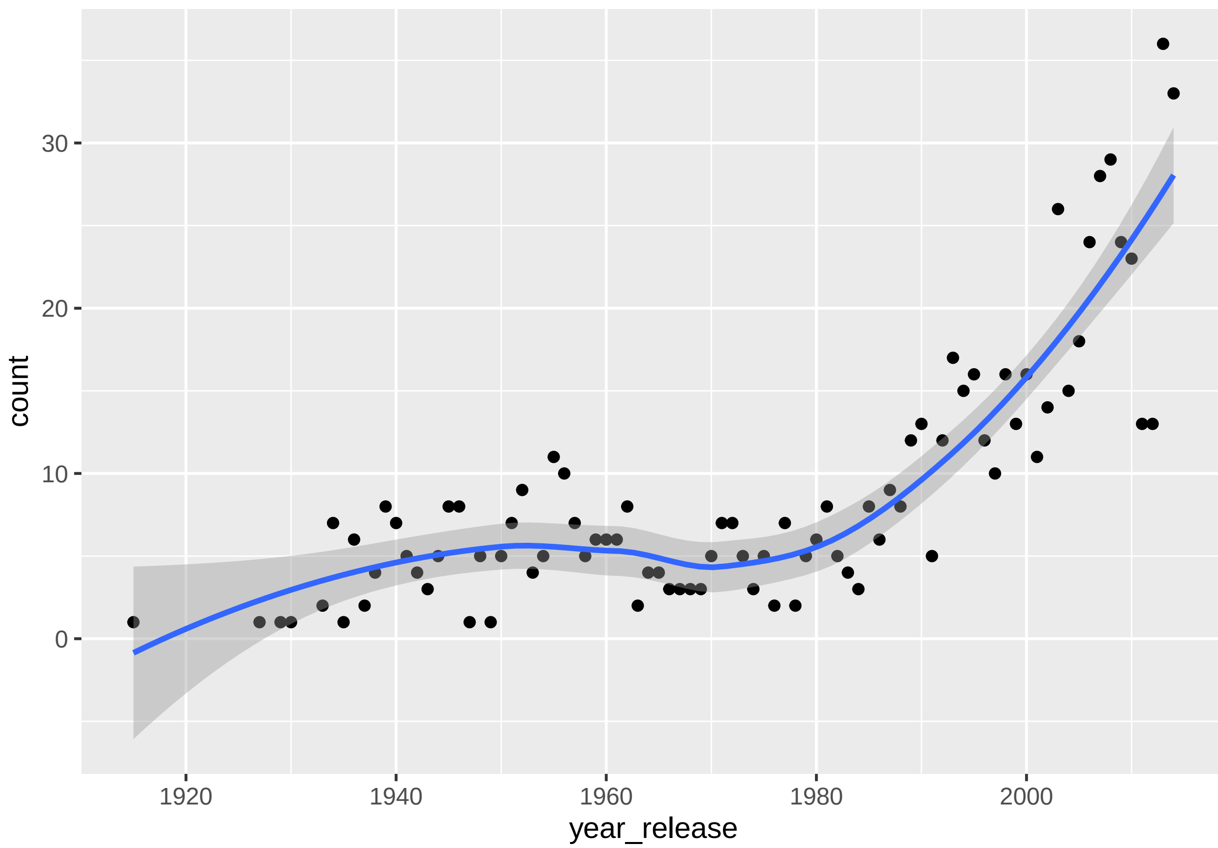
```r
head(biopics)
```

```
## # A tibble: 6 x 14
##   title site  country year_release box_office director number_of_subje~
##   <chr> <chr> <chr>          <int>      <dbl> <chr>               <int>
## 1 10 R~ tt00~ UK              1971         NA Richard~                1
## 2 12 Y~ tt20~ US/UK           2013   56700000 Steve M~                1
## 3 127 ~ tt15~ US/UK           2010   18300000 Danny B~                1
## 4 1987  tt28~ Canada          2014         NA Ricardo~                1
## 5 20 D~ tt01~ US              1998     537000 Myles B~                1
## 6 21    tt04~ US              2008   81200000 Robert ~                1
## # ... with 7 more variables: subject <chr>, type_of_subject <chr>,
## #   race_known <chr>, subject_race <chr>, person_of_color <lgl>,
## #   subject_sex <chr>, lead_actor_actress <chr>
```

(a)

```r
a <- biopics %>% group_by(year_release) %>% summarise(count = n())
a <- as.data.frame(a)
ggplot(a, aes(x = year_release, y = count)) +
  geom_point() + geom_smooth(method = 'auto')
```

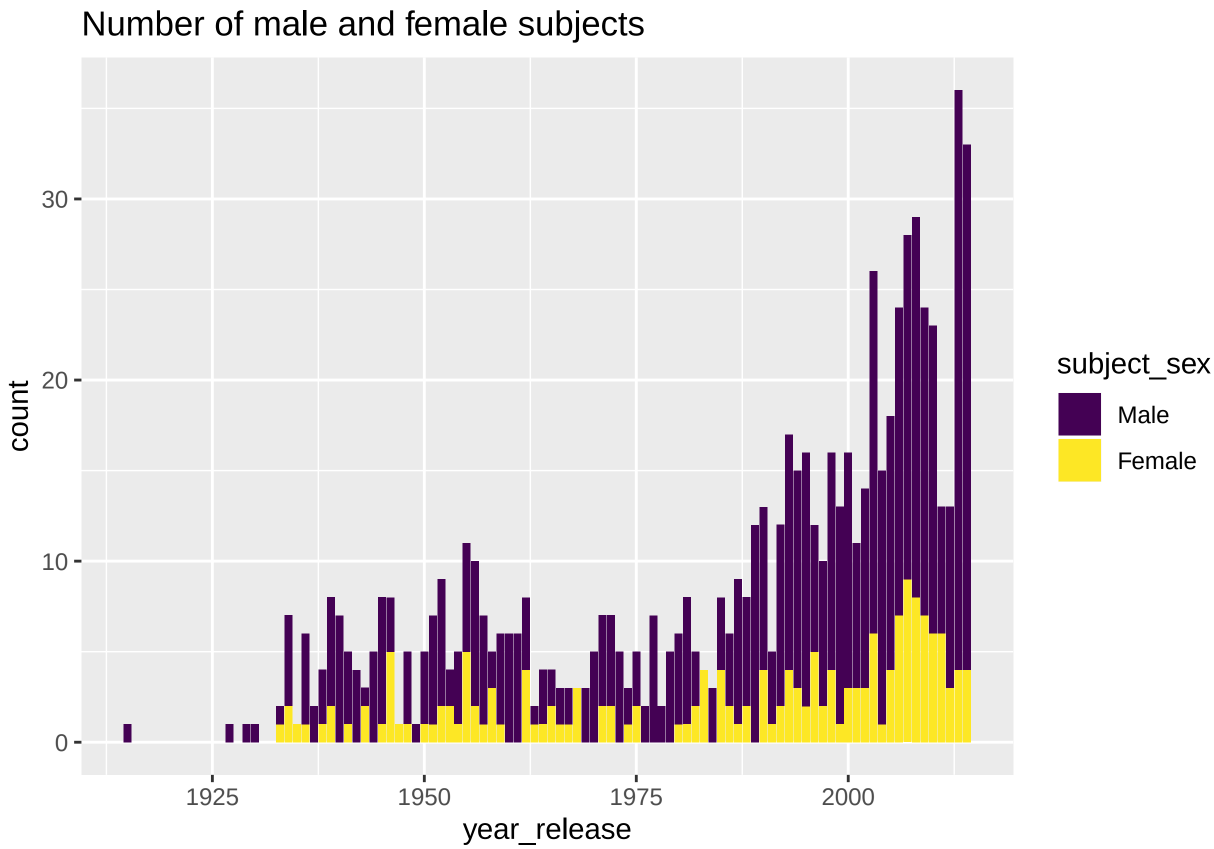
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Clearly, from the above plot, the total number of biopics released per year has increased over time.

(b)

```r
b = biopics %>% mutate(subject_sex = fct_infreq(subject_sex))
ggplot(b,aes(x=year_release,fill=subject_sex)) +
geom_bar() + scale_fill_viridis_d() + labs(title="Number of male and female subjects")
```




```r
head(b)
```

```
## # A tibble: 6 x 14
##   title site  country year_release box_office director number_of_subje~
##   <chr> <chr> <chr>          <int>      <dbl> <chr>               <int>
## 1 10 R~ tt00~ UK              1971         NA Richard~                1
## 2 12 Y~ tt20~ US/UK           2013   56700000 Steve M~                1
## 3 127 ~ tt15~ US/UK           2010   18300000 Danny B~                1
## 4 1987  tt28~ Canada          2014         NA Ricardo~                1
## 5 20 D~ tt01~ US              1998     537000 Myles B~                1
## 6 21    tt04~ US              2008   81200000 Robert ~                1
## # ... with 7 more variables: subject <chr>, type_of_subject <chr>,
## #   race_known <chr>, subject_race <chr>, person_of_color <lgl>,
## #   subject_sex <fct>, lead_actor_actress <chr>
```

(c)

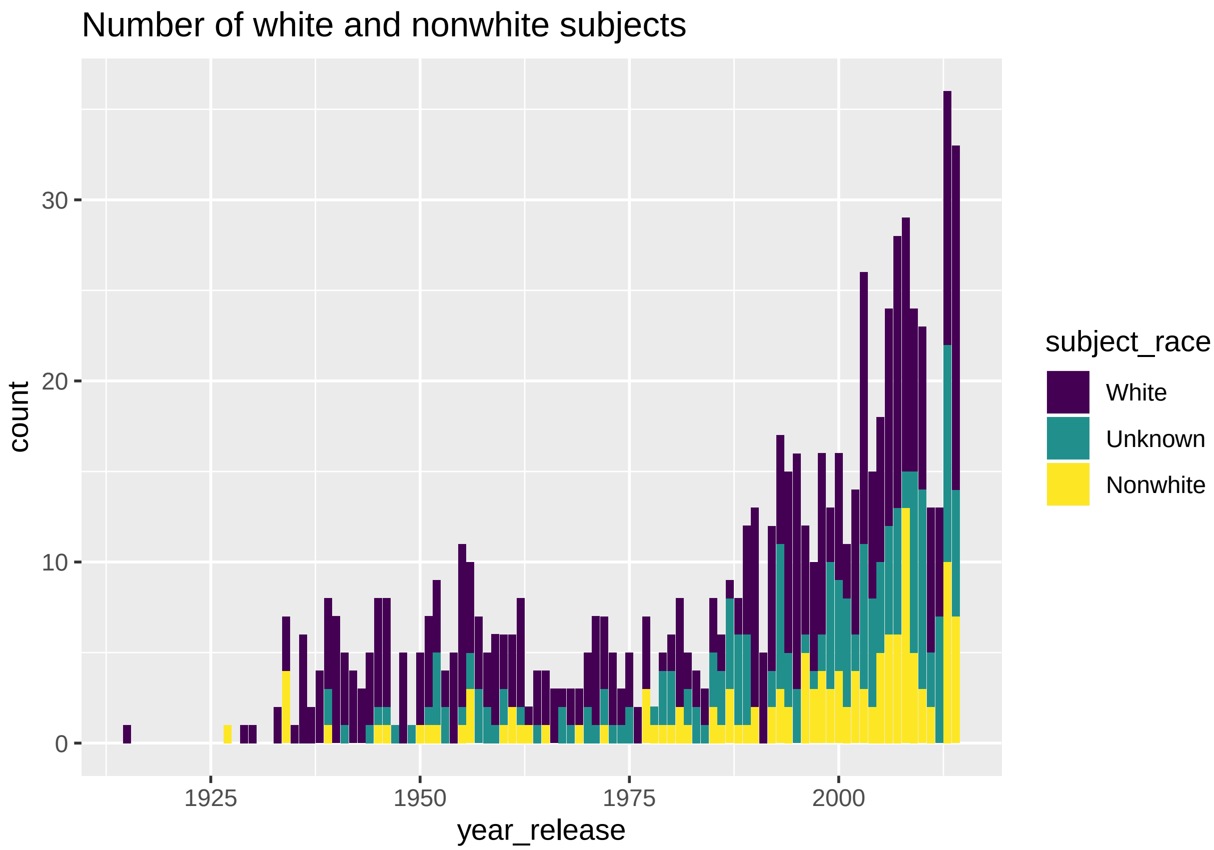
Recategorizing subject race into "White", "Nonwhite", and "Unknown":

```r
biopics$subject_race[is.na(biopics$subject_race)] <- "Unknown"
biopics$subject_race[biopics$subject_race != "White" &
                       biopics$subject_race != "Unknown"] <- "Nonwhite"
```

```r
head(biopics$subject_race)
```

```
## [1] "Unknown"  "Nonwhite" "Unknown"  "White"    "Unknown"  "Nonwhite"
```

```
c = biopics %>% mutate(subject_race = fct_infreq(subject_race))
ggplot(c,aes(x=year_release,fill=subject_race)) +
geom_bar() + scale_fill_viridis_d() + labs(title="Number of white and nonwhite subjects")
```



(d)

```
c$years[(c$year_release >= 1915) & (c$year_release < 1934)] <- "1915~1934"
```

```
## Warning: Unknown or uninitialised column: 'years'.
```

```
c$years[(c$year_release >= 1934) & (c$year_release < 1954)] <- "1934~1954"
c$years[(c$year_release >= 1954) & (c$year_release < 1974)] <- "1954~1974"
c$years[(c$year_release >= 1974) & (c$year_release < 1994)] <- "1974~1994"
c$years[(c$year_release >= 1994) & (c$year_release <= 2014)] <- "1994~2014"
```

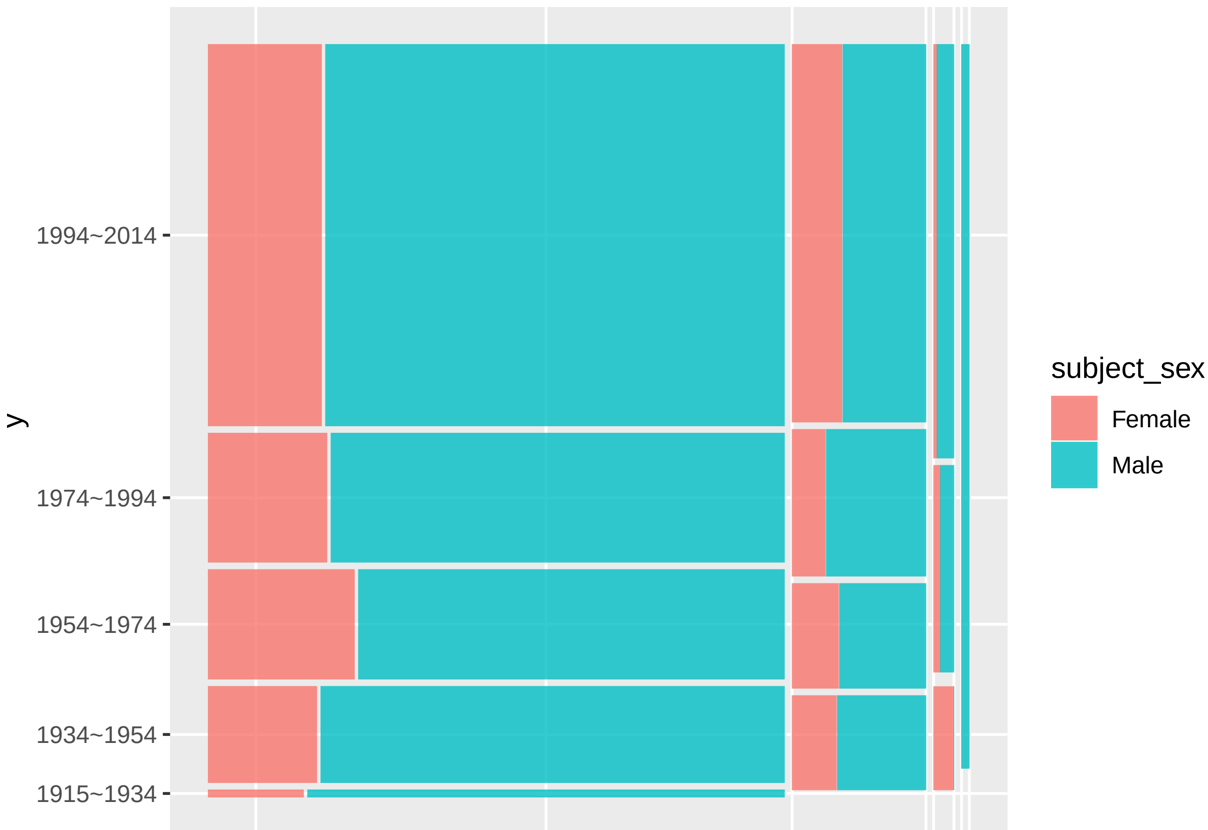
```
head(c$years)
```

```
## [1] "1954~1974" "1994~2014" "1994~2014" "1994~2014" "1994~2014" "1994~2014"
```

```
ggplot(c) + geom_mosaic(aes(x=product(years, number_of_subjects),
                                fill=subject_race)) +
theme(axis.title.x =element_blank(),axis.text.x =element_blank(),
axis.ticks.x=element_blank())
```

From the above plot, we can see that the "Nonwhite" group is the most underrepresented in biopics based on number of subjects since the blue areas are relatively smaller than the other two.

```
ggplot(c) + geom_mosaic(aes(x=product(years, number_of_subjects),
                                fill=subject_sex)) +
theme(axis.title.x =element_blank(),axis.text.x =element_blank(),
axis.ticks.x=element_blank())
```



From the above plot, we can see that the "Female" group is the most underrepresented in biopics based on number of subjects since the the total red areas are relatively smaller than the blue areas.

(e)

```
d1 <- c %>% group_by(year_release,subject_race) %>%
  summarise(count=n()) %>% mutate(prop=count/sum(count))
d1
```

```
## # A tibble: 192 x 4
## # Groups:   year_release [86]
##    year_release subject_race count  prop
##           <int> <fct>        <int> <dbl>
##  1         1915 White            1 1
##  2         1927 Nonwhite         1 1
##  3         1929 White            1 1
##  4         1930 White            1 1
##  5         1933 White            2 1
##  6         1934 White            3 0.429
##  7         1934 Nonwhite         4 0.571
##  8         1935 White            1 1
##  9         1936 White            6 1
## 10         1937 White            2 1
## # ... with 182 more rows
```

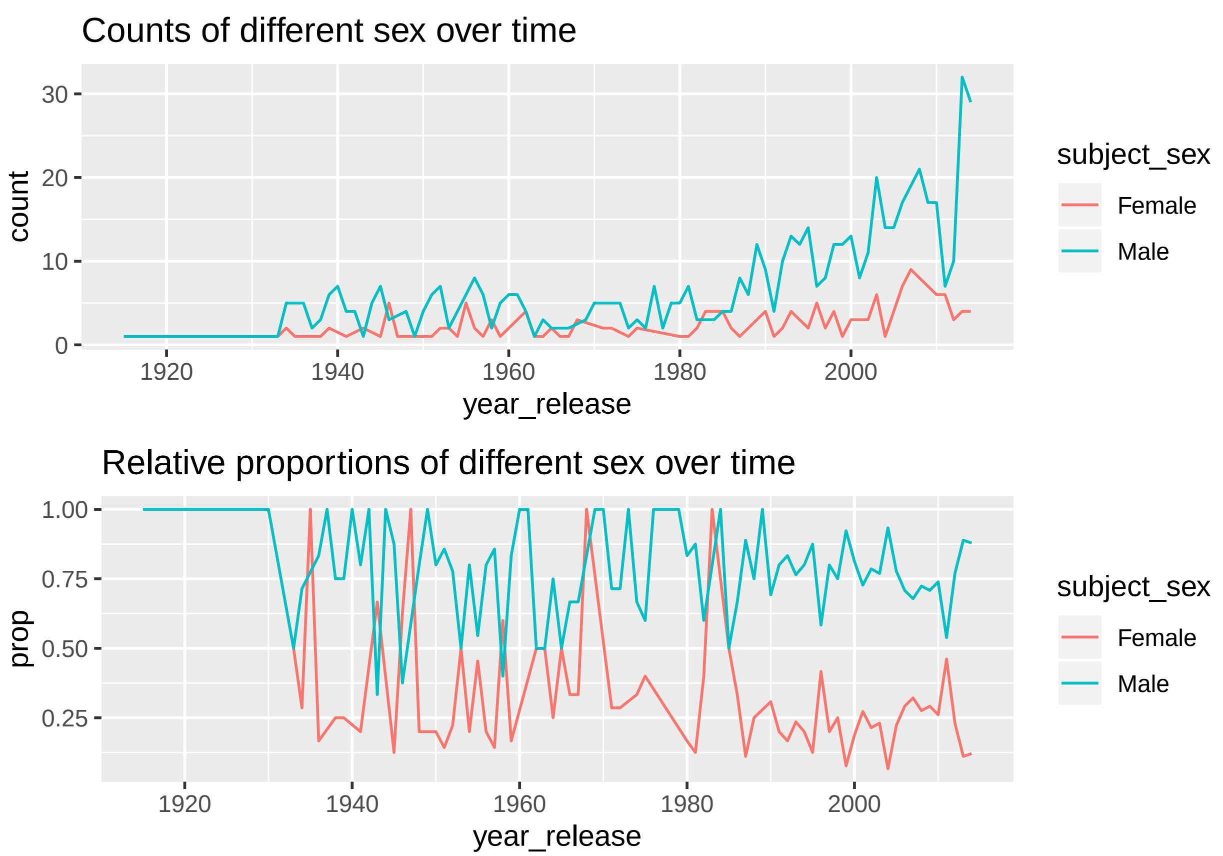
```
d2 <- c %>% group_by(year_release,subject_sex) %>%
  summarise(count=n()) %>% mutate(prop=count/sum(count))
d2
```

```
## # A tibble: 148 x 4
```

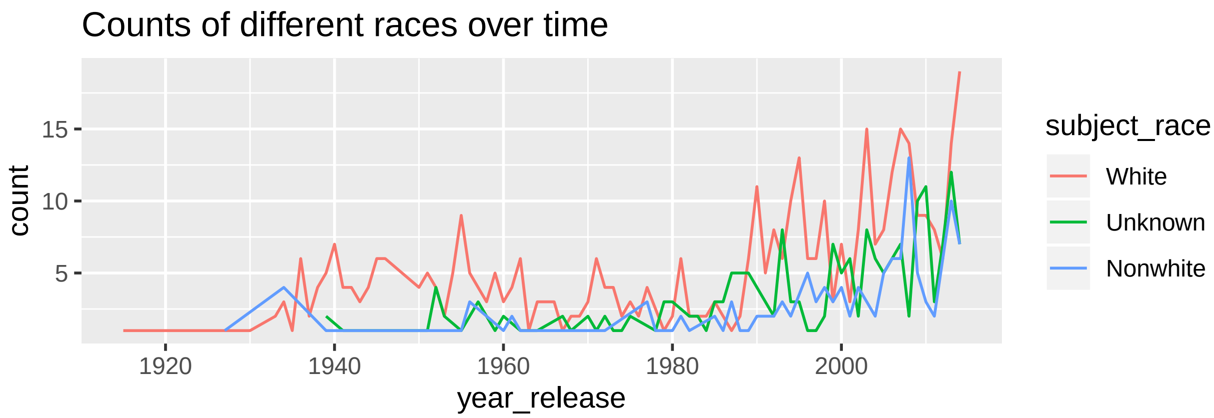
```
## # Groups:   year_release [86]
##    year_release subject_sex count  prop
##           <int> <chr>       <int> <dbl>
##  1         1915 Male            1 1
##  2         1927 Male            1 1
##  3         1929 Male            1 1
##  4         1930 Male            1 1
##  5         1933 Female          1 0.5
##  6         1933 Male            1 0.5
##  7         1934 Female          2 0.286
##  8         1934 Male            5 0.714
##  9         1935 Female          1 1
## 10         1936 Female          1 0.167
## # ... with 138 more rows
```

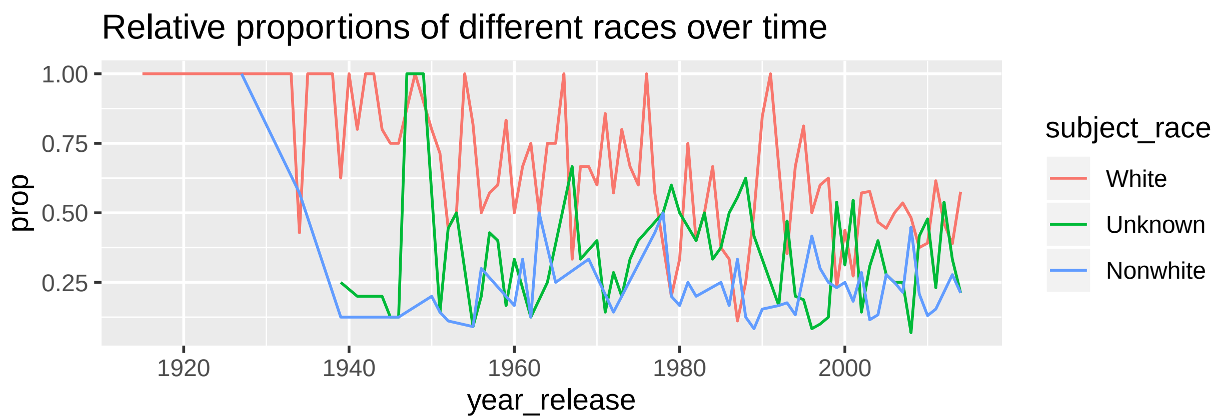
(f)

```
p1 = ggplot(d2, aes(year_release, y=count)) + geom_line(aes(colour=subject_sex)) +
  labs(title = "Counts of different sex over time")
p2 = ggplot(d2, aes(year_release, y=prop)) + geom_line(aes(colour=subject_sex)) +
  labs(title = "Relative proportions of different sex over time")
grid.arrange(p1,p2)
```




```
p3 = ggplot(d1, aes(year_release, y=count)) + geom_line(aes(colour=subject_race)) +
  labs(title = "Counts of different races over time")
p4 = ggplot(d1, aes(year_release, y=prop)) + geom_line(aes(colour=subject_race)) +
  labs(title = "Relative proportions of different races over time")
grid.arrange(p3,p4)
```





Based on the above plots, we conclude that although the imbalance of different races is improving over the time, it is still a huge problem when it comes to different sex. The line plot showing the counts of different sex and races over time indicates that the number of different races increases in a more and more consistent pace while the number of the male subjects has a more significant growth than that of the female subjects as time passes by. Moreover, from the line plot showing the relative proportions of subjects over time, the lines of different races tend to fluctuate within a smaller and smaller range while the line representing the relative proportions of male subjects stays at the top of the plot almost all the time. The above reasons all lead to the conclusion stated previously.

Question 2

```r
library(heplots)
data(Diabetes)
```

Data exploration

```r
summary(Diabetes)
```

```
##      relwt           glufast        glutest          instest     
##  Min.   :0.7100   Min.   : 70   Min.   : 269.0   Min.   : 10.0  
##  1st Qu.:0.8800   1st Qu.: 90   1st Qu.: 352.0   1st Qu.:118.0  
##  Median :0.9800   Median : 97   Median : 413.0   Median :156.0  
##  Mean   :0.9773   Mean   :122   Mean   : 543.6   Mean   :186.1  
##  3rd Qu.:1.0800   3rd Qu.:112   3rd Qu.: 558.0   3rd Qu.:221.0  
##  Max.   :1.2000   Max.   :353   Max.   :1568.0   Max.   :748.0  
##       sspg                       group   
##  Min.   : 29.0   Normal           :76
```

```
##   1st Qu.:100.0   Chemical_Diabetic:36
##   Median :159.0   Overt_Diabetic   :33
##   Mean   :184.2
##   3rd Qu.:257.0
##   Max.   :480.0
```

```r
head(Diabetes)
```

```
##   relwt glufast glutest instest sspg  group
## 1  0.81      80     356     124   55 Normal
## 2  0.95      97     289     117   76 Normal
## 3  0.94     105     319     143  105 Normal
## 4  1.04      90     356     199  108 Normal
## 5  1.00      90     323     240  143 Normal
## 6  0.76      86     381     157  165 Normal
```

(a)

```r
Diabetes %>% group_by(group) %>% summarise_all(list(Avg=mean,Med=median)) %>%
  pivot_longer(cols=c("relwt_Avg", "relwt_Med", "glufast_Avg", "glufast_Med",
                      "glutest_Avg", "glutest_Med", "instest_Avg",
                      "instest_Med", "sspg_Avg", "sspg_Med"),names_to = "Measure") %>%
  pivot_wider(id_cols=Measure,names_from=group) %>% arrange(desc(Measure))
```

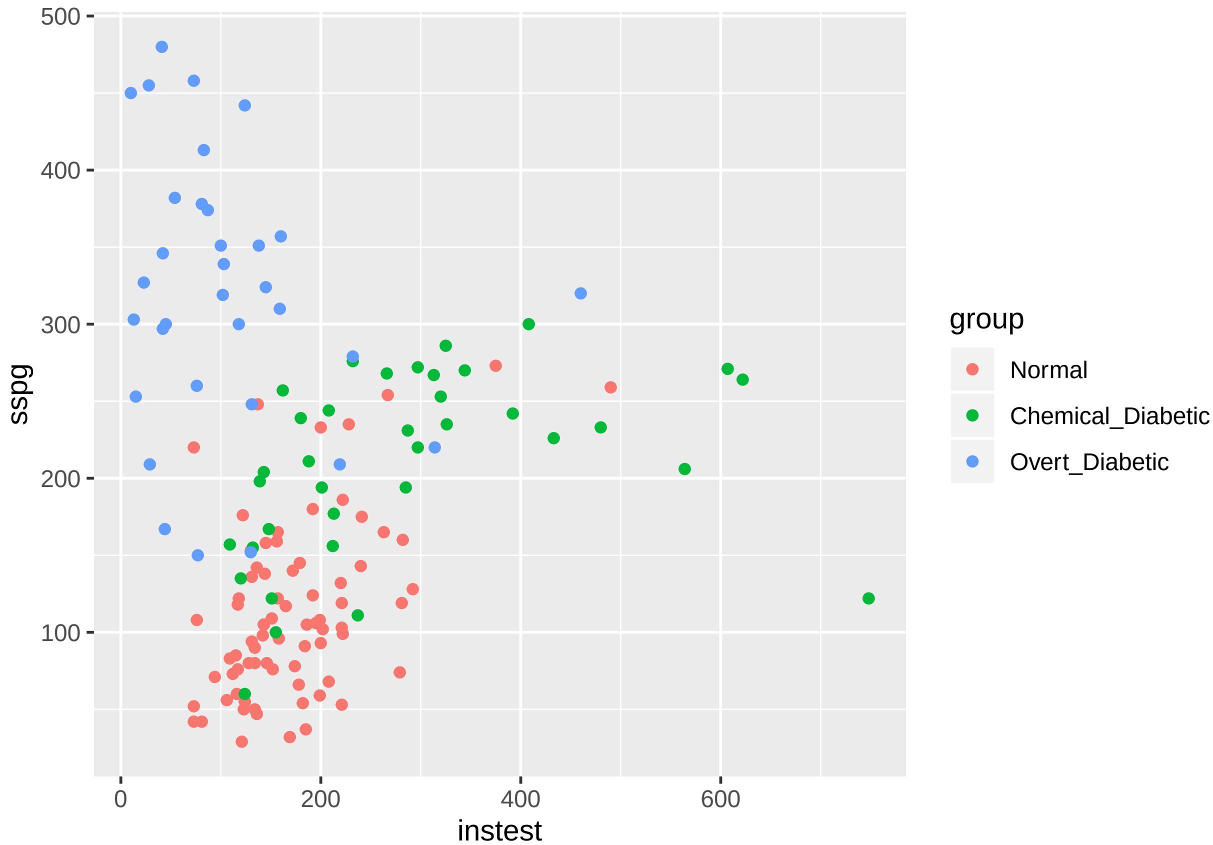
```
## # A tibble: 10 x 4
##    Measure      Normal Chemical_Diabetic Overt_Diabetic
##    <chr>         <dbl>             <dbl>          <dbl>
##  1 sspg_Med    105               223            320
##  2 sspg_Avg    114               209.           319.
##  3 relwt_Med     0.95              1.06           0.98
##  4 relwt_Avg     0.937             1.06           0.984
##  5 instest_Med 157               252.            83
##  6 instest_Avg 173.              288            106
##  7 glutest_Med 353               476.           972
##  8 glutest_Avg 350.              494.          1044.
##  9 glufast_Med  90                99.5          203
## 10 glufast_Avg  91.2              99.3          218.
```

From the table above, variable "sspg", "instest", and "glutest" seem to differentiate amongst the different types of diabetes very well.

(b)

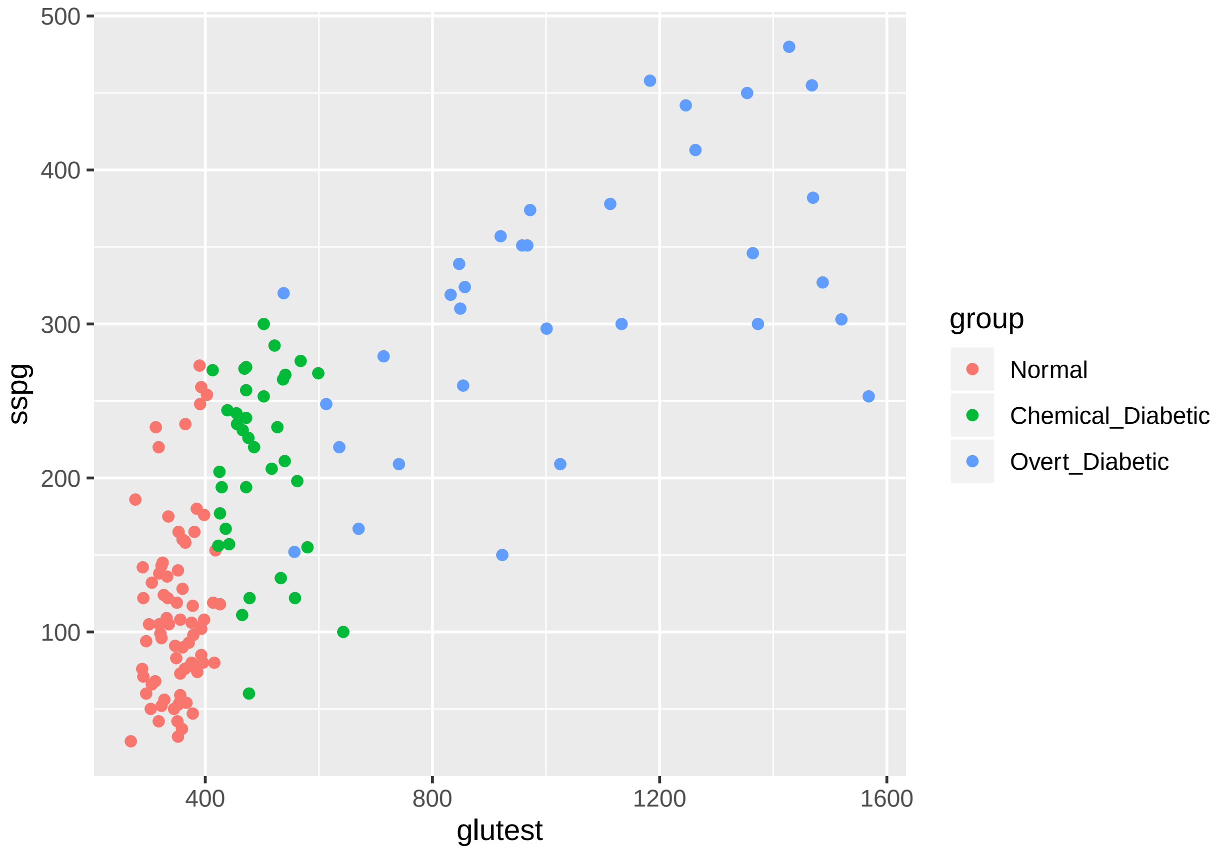
"sspg" versus "insulin test":

```r
ggplot(Diabetes, aes(x = instest, y = sspg, colour = group)) + geom_point()
```



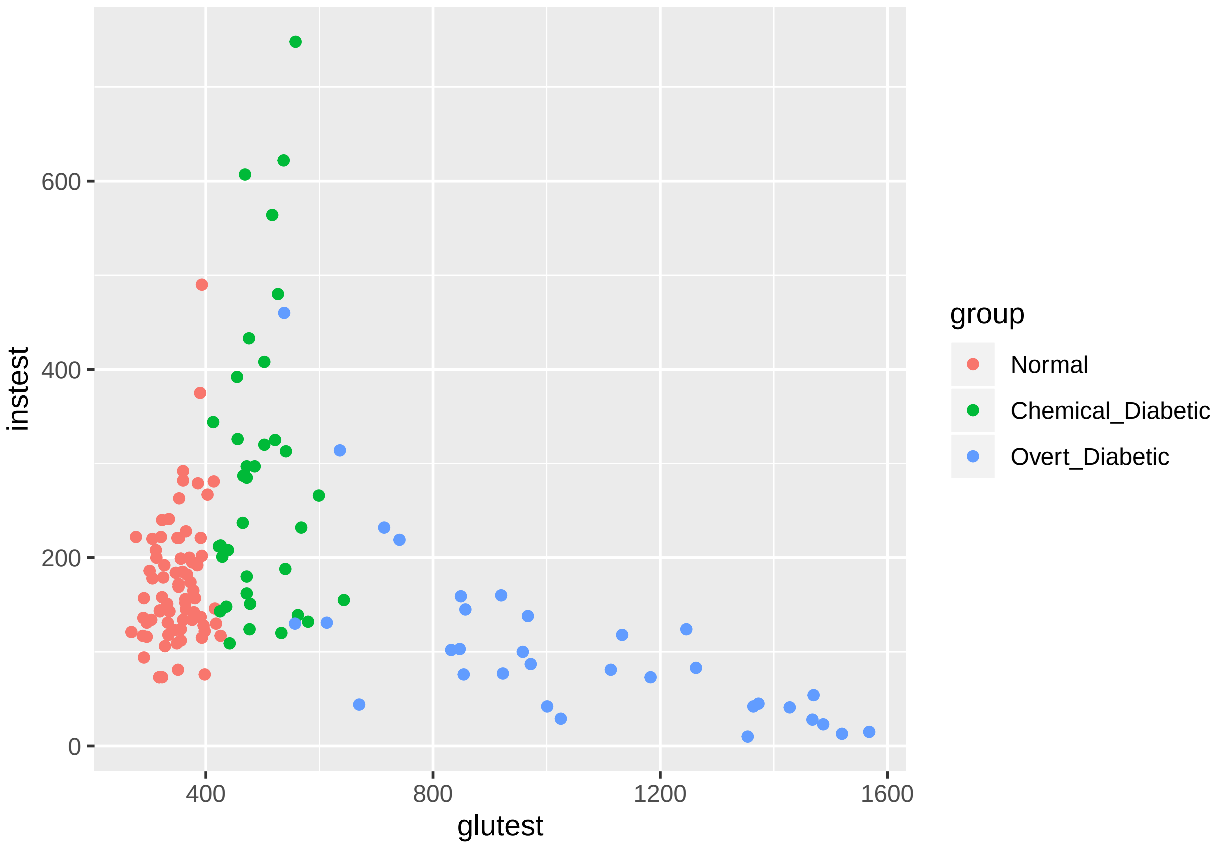
"sspg" versus "glucose test":

```
ggplot(Diabetes, aes(x = glutest, y = sspg, colour = group)) + geom_point()
```



“insulin test” versus “glucose test”:

```
ggplot(Diabetes, aes(x = glutest, y = instest, colour = group)) + geom_point()
```



"

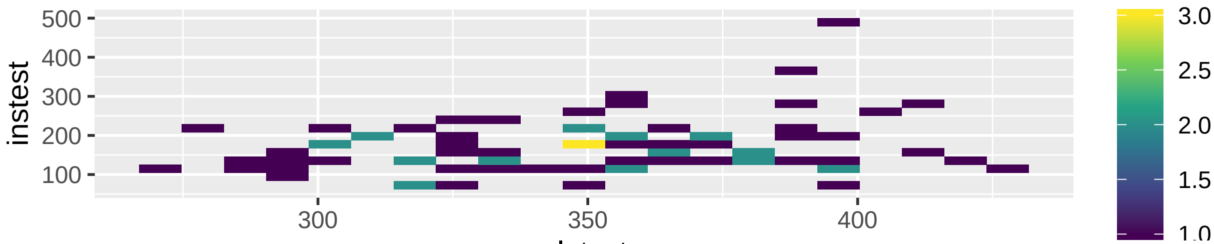
The pair "“insulin test” and “glucose test” seems to allow for the strongest distinction amongst the three groups.

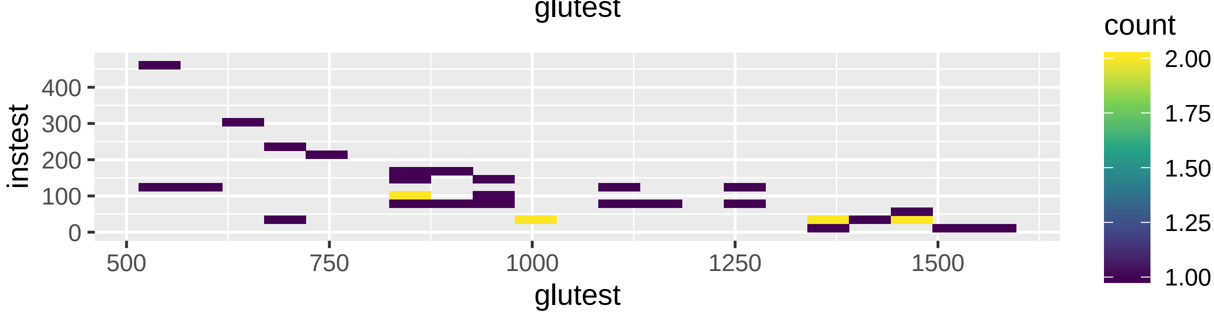
(c)

Histograms:

```
normal <- Diabetes %>% filter(group == "Normal")
chemical <- Diabetes %>% filter(group == "Chemical_Diabetic")
overt <- Diabetes %>% filter(group == "Overt_Diabetic")
```

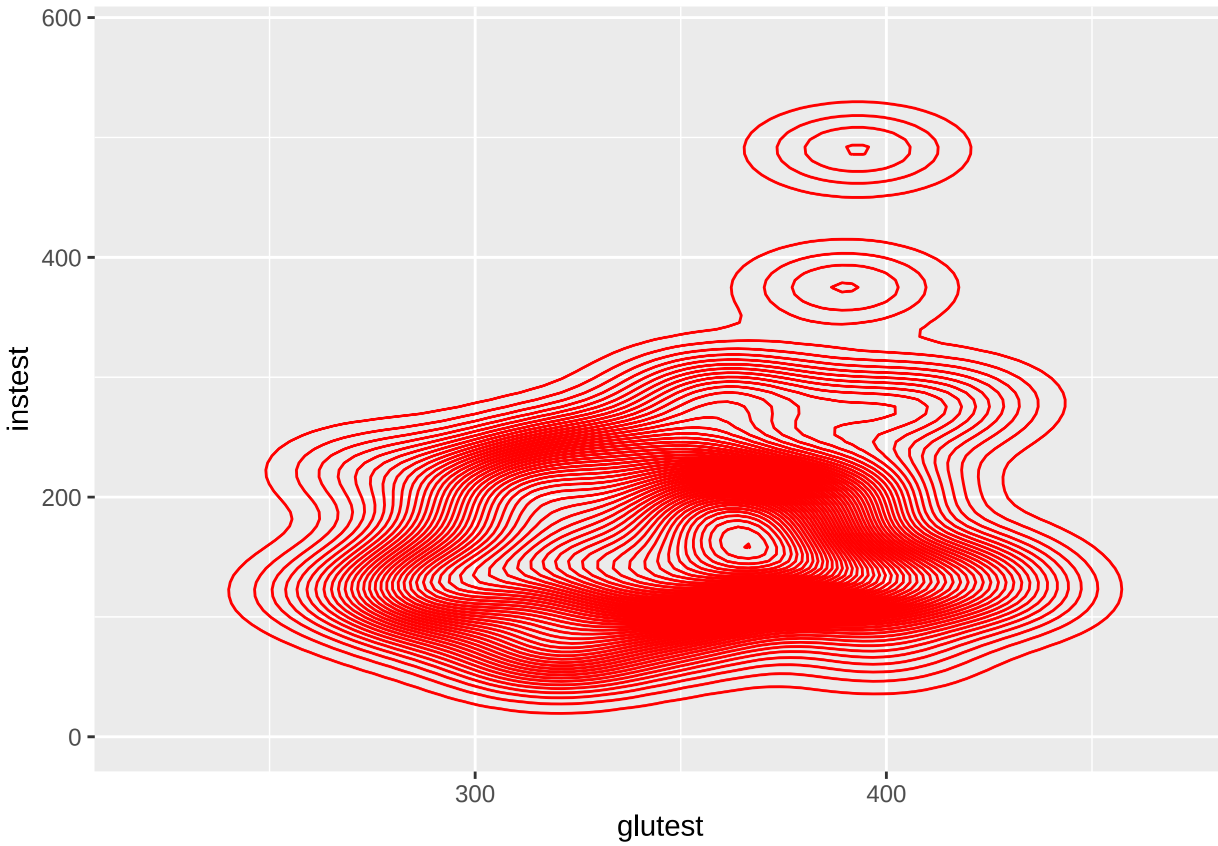
```
p = ggplot(normal,aes(x=glutest,y=instest)) + geom_bin2d(bins=20) +
  scale_fill_continuous(type = "viridis") + labs(x="glutest",y="instest")
q = ggplot(chemical,aes(x=glutest,y=instest)) + geom_bin2d(bins=20) +
  scale_fill_continuous(type = "viridis") + labs(x="glutest",y="instest")
n = ggplot(overt,aes(x=glutest,y=instest)) + geom_bin2d(bins=20) +
  scale_fill_continuous(type = "viridis") + labs(x="glutest",y="instest")
grid.arrange(p,q,n)
```



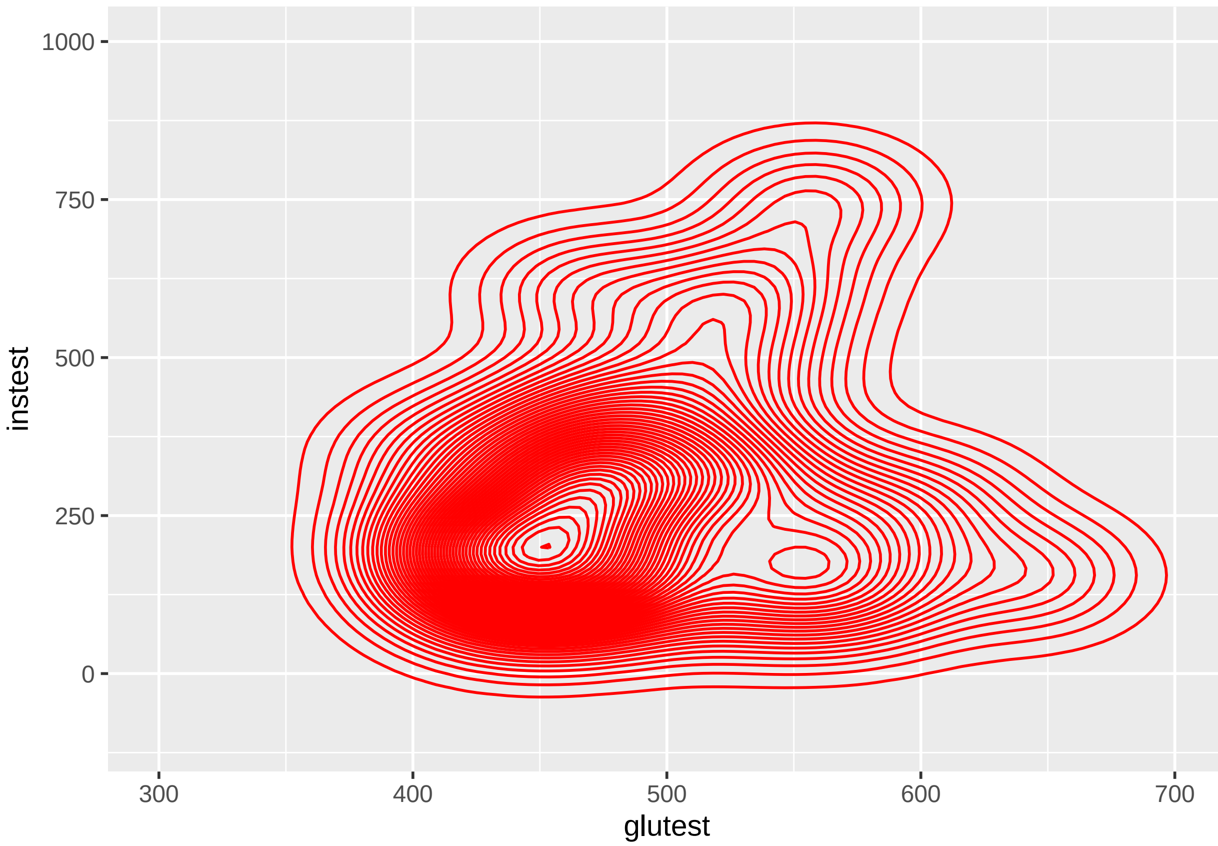


Contour plots:

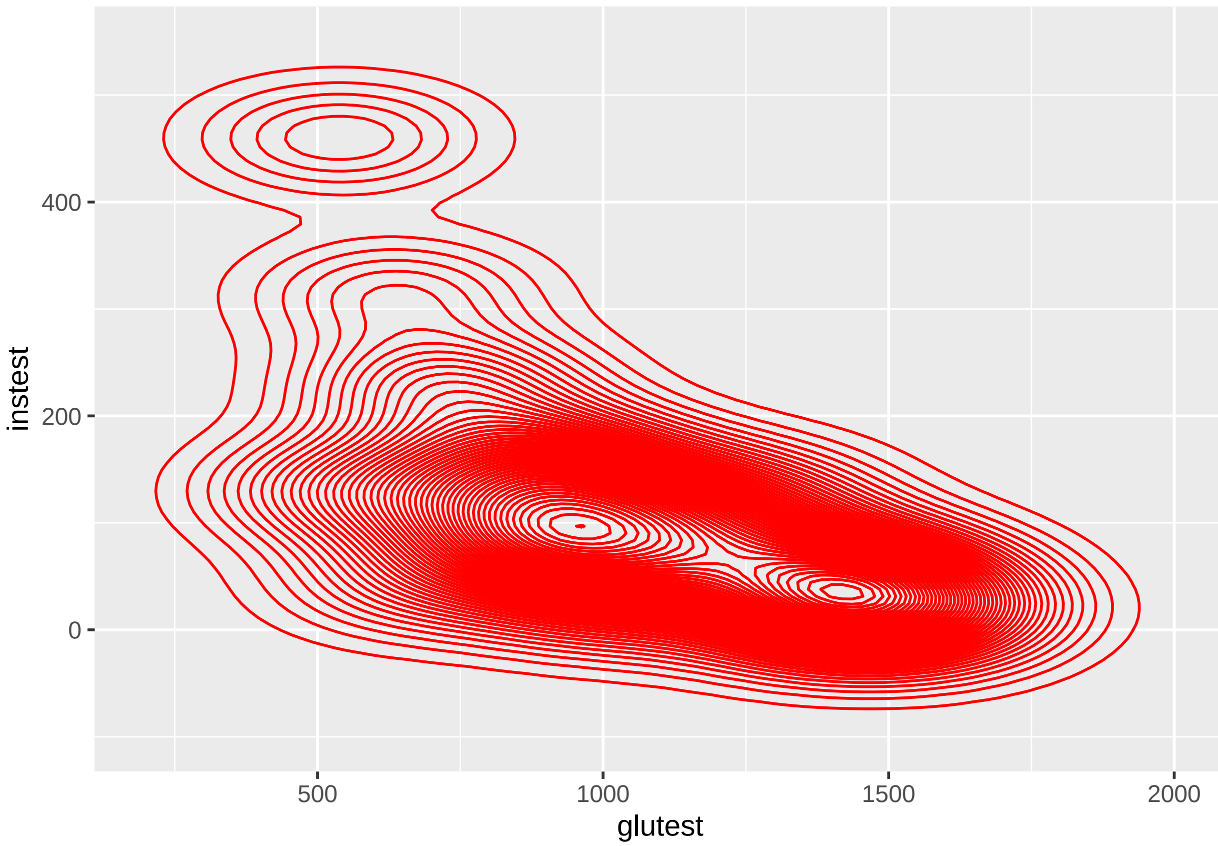
```
ggplot(normal,aes(x=glutest,y=instest)) + geom_density_2d(col="red",bins=50) +
  labs(x="glutest",y="instest") +
  ylim(c(0,580)) + xlim(c(220,470))
```




```r
ggplot(chemical,aes(x=glutest,y=instest)) + geom_density_2d(col="red",bins=50) +
  labs(x="glutest",y="instest") +
  ylim(c(-100,1000)) + xlim(c(300,700))
```




```
ggplot(overt,aes(x=glutest,y=instest)) + geom_density_2d(col="red",bins=50) +
  labs(x="glutest",y="instest") +
  ylim(c(-100,550)) + xlim(c(200,2000))
```



These plots do provide useful summaries of the differences in distributions in the three groups. From the previous visualization, we picked out the subspace "instest"×"glutest" which "allows for the strongest distinction amongst the three groups". That is to say, these marginal distributions best describe the corresponding original distributions. Also, the histograms and contour plots gives information about the mode, mean, variance and other properties of the distributions. So the differences in distributions in the three groups are clearly visualized.