Math208-A2

MATH 208 - Assignment 2 - Christopher Zheng - 206760794

Question 1:

```
#install.packages("fivethirtyeight")
library(fivethirtyeight)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------
------------------------------------------------------------------------------------------ t
idyverse 1.2.1 --
```

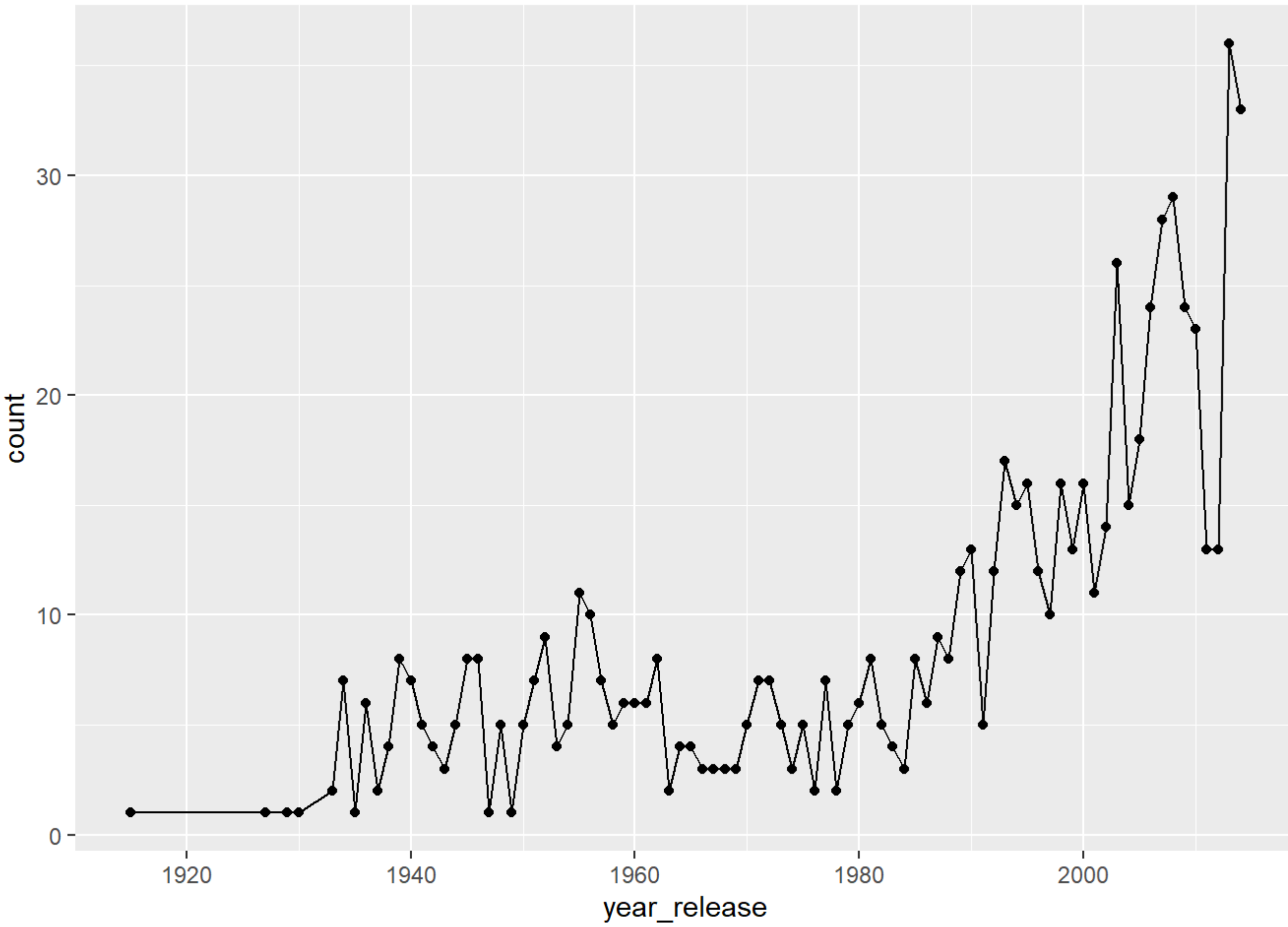
```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
library(ggplot2)

df <- tbl_df(biopics)
```

a. Using the plot of your choice, assess whether the total number of biopics released per year has increased over time based on the data collected from the IMDB movie database.

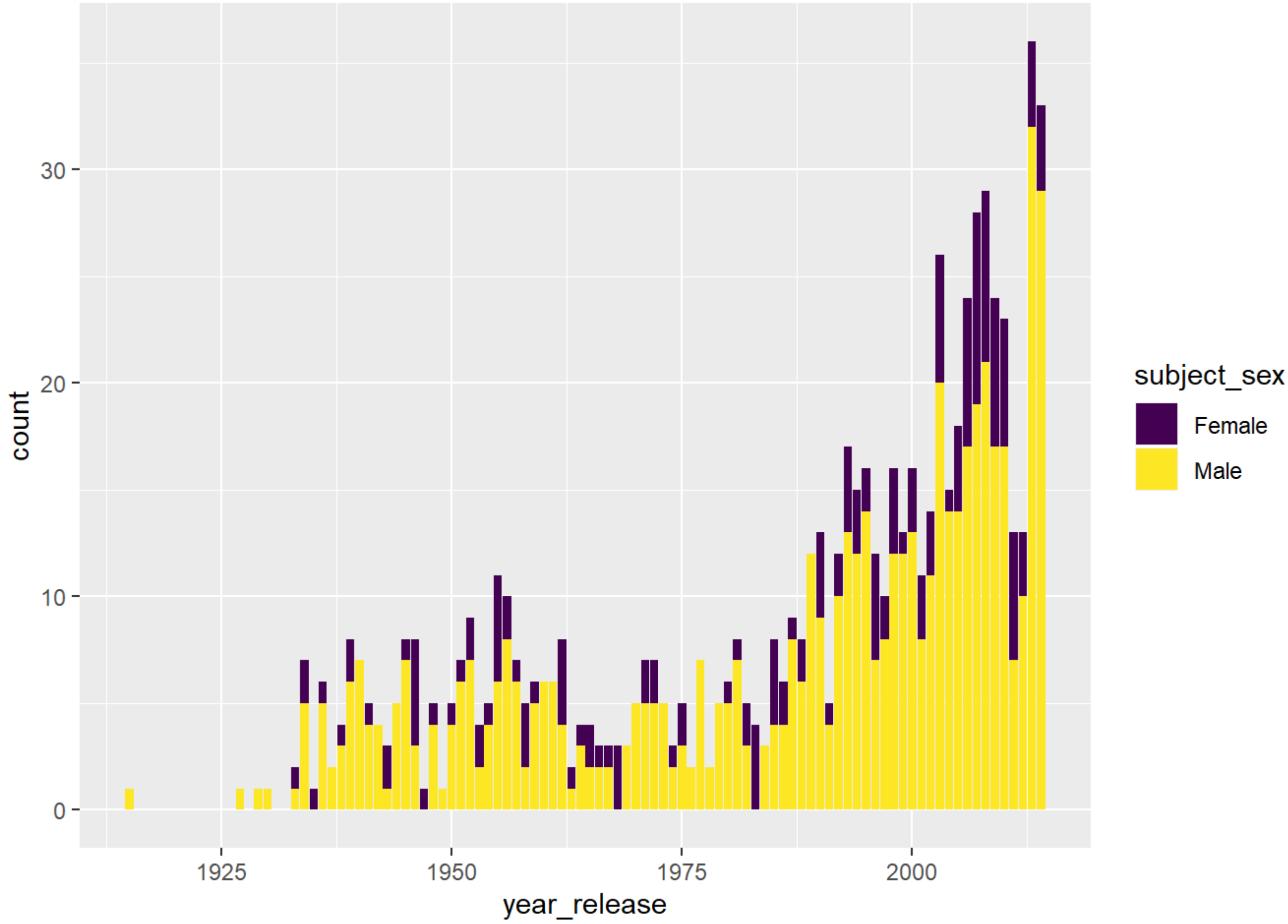
```
df_per_year <- df %>% group_by(year_release) %>% summarise(count=n())

ggplot(df_per_year,aes(x=year_release,y=count,group=1)) +
  geom_point() + geom_line(stat="summary",fun.y=mean)
```




```
#In conlcusion, the total number of biopics released per year has increased over time based on t
he timeline.
```

b. Produce a stacked barplot similar to the barplot in the original article showing the relative numbers of male and female subjects over time (Note the figures will not exactly be the same as the data in the article figures is not the same as in the dataset).

```
df_per_year <- df %>% group_by(year_release) %>% mutate(count=n())

ggplot(df_per_year,aes(x=year_release,fill=subject_sex)) +
  geom_bar() + scale_fill_viridis_d()
```

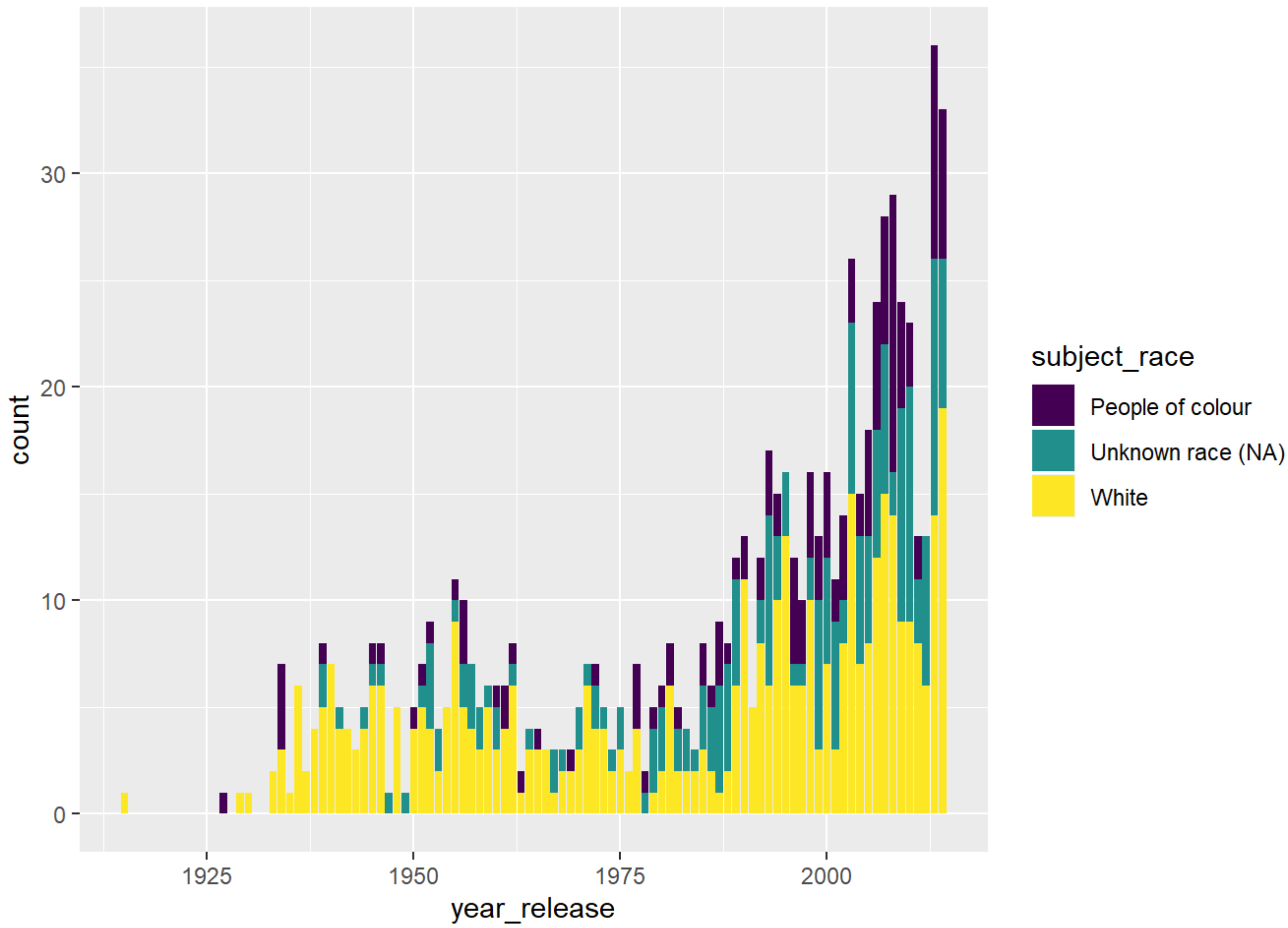



```
head(df_per_year)
```

```
## # A tibble: 6 x 15
## # Groups:   year_release [6]
##   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 8 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>, count <int>
```

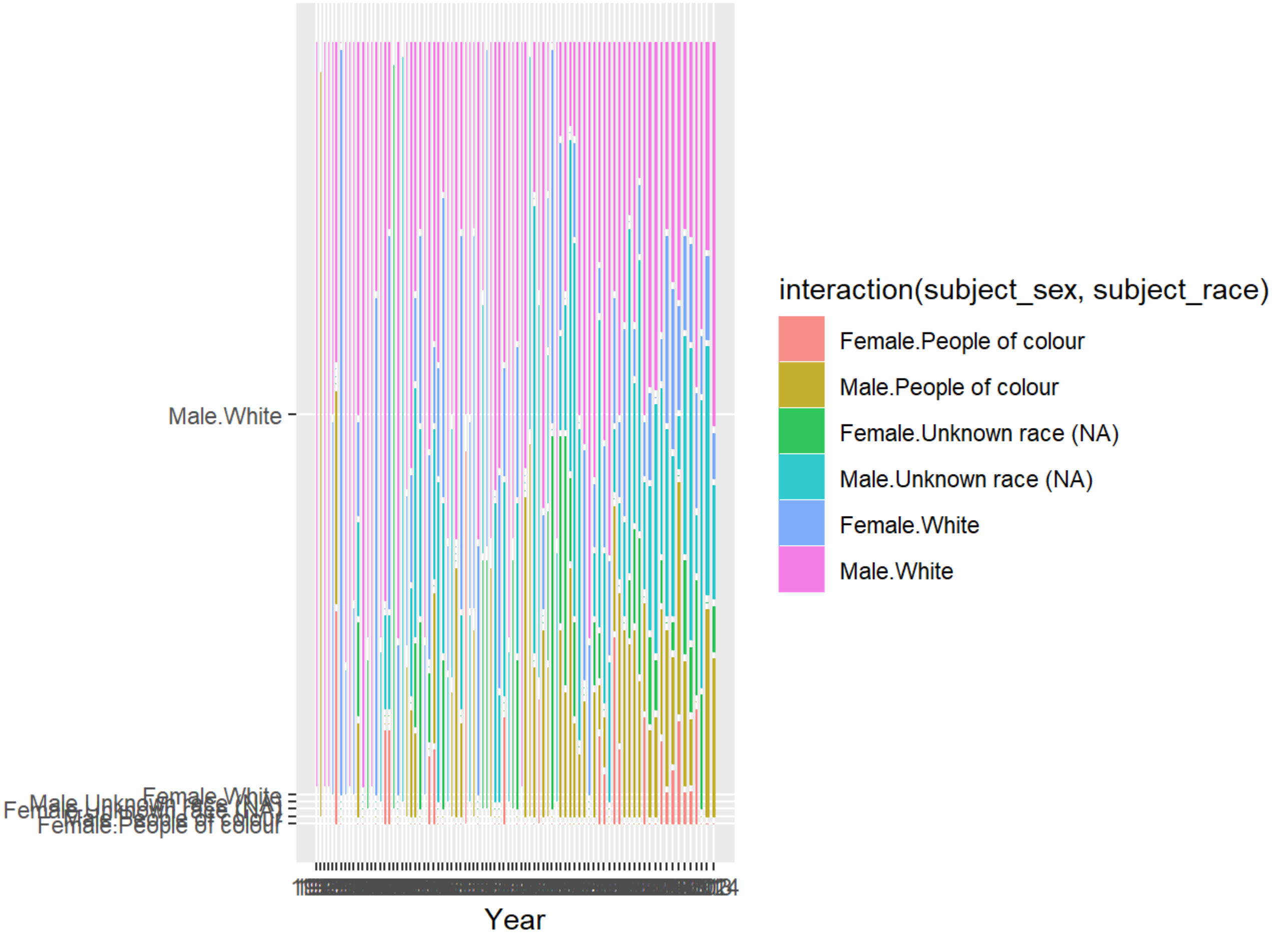
c. Produce a stacked barplot similar to the barplot in the original article showing the relative numbers of white subjects, subjects who are persons of color, and unknown race subjects over time. (Mote the figures will not exactly be the same as the data in the article figures is not the same as in the dataset).

```
df_per_year <- df %>% group_by(year_release) %>% mutate(count=n()) %>%
  mutate(subject_race = ifelse(is.na(subject_race),"Unknown race (NA)",ifelse(subject_race=="Whi
te",subject_race,"People of colour")) )
#df_per_year
ggplot(df_per_year,aes(x=year_release,fill=subject_race)) +
  geom_bar() + scale_fill_viridis_d()
```



d. Based on a mosaic plot (collapsing over year of release), which sex / white-nonwhite-NA group is the most underrepresented in biopics based on number of subjets?

```
library(ggmosaic)
df_per_year <- df %>% group_by(year_release) %>% mutate(count=n()) %>%
  mutate(subject_race = ifelse(is.na(subject_race),"Unknown race (NA)",ifelse(subject_race=="Whi
te",subject_race,"People of colour")) )


ggplot(df_per_year) +
  geom_mosaic(aes(x=product(year_release), fill=interaction(subject_sex,subject_race)))+
  xlab("Year") + ylab("")
```




```
# Based on the mosaic plot, females who are people of colours are the most underrepresented.
```

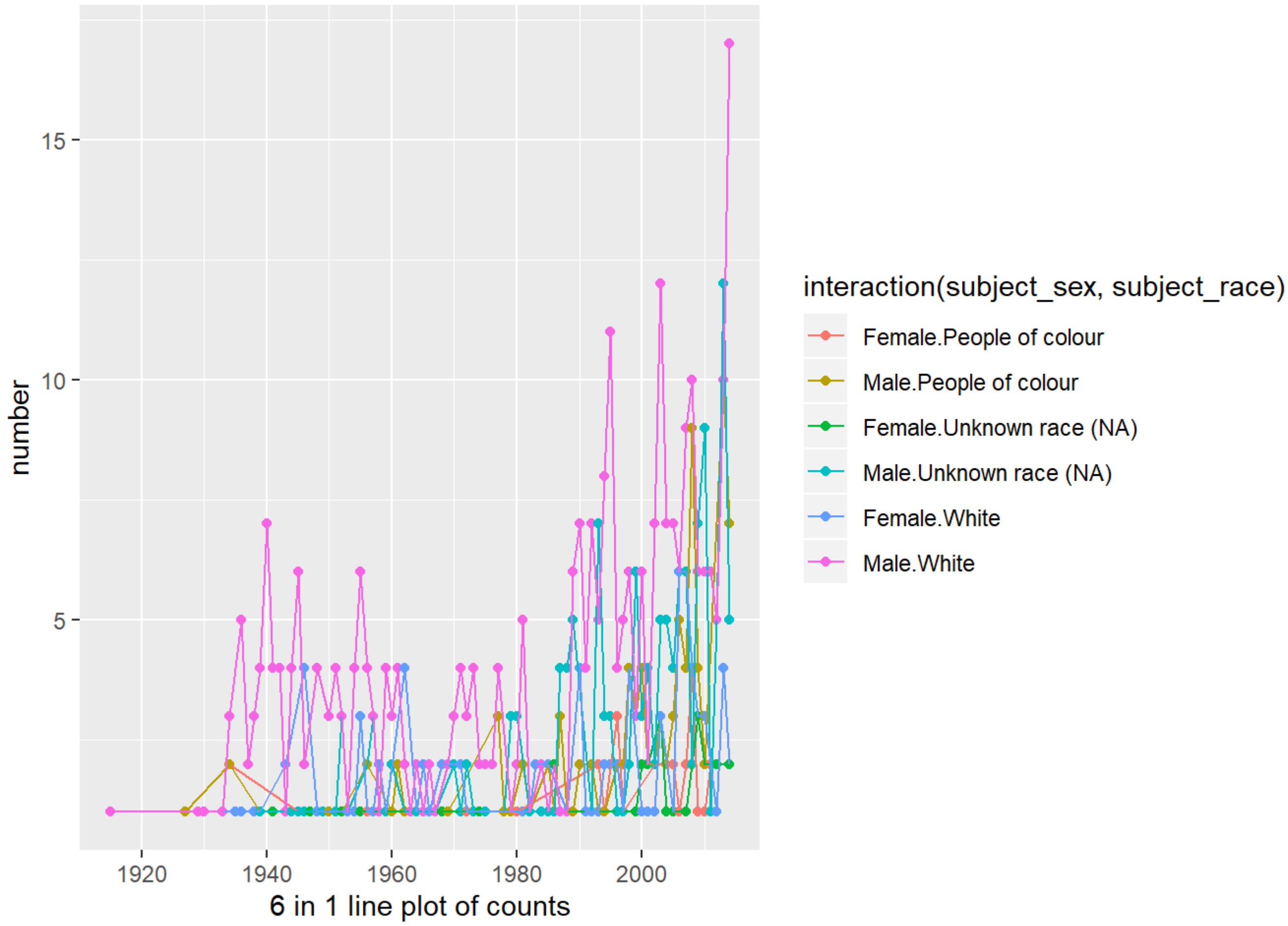
e. Produce a summary table containing counts and proportions of biopic subjects per year for each sex/white-nonwehite-NA factor combination.

```
library(dplyr)

df_per_year <- df %>% group_by(year_release) %>% mutate(count=n()) %>%
  mutate(subject_race = ifelse(is.na(subject_race),"Unknown race (NA)",ifelse(subject_race=="Whi
te",subject_race,"People of colour")) ) %>%
  select(year_release,subject_race,subject_sex,count)
#df_per_year
df_summary <- df_per_year %>% group_by(year_release,subject_sex,subject_race) %>% mutate(number
 = n(),proportion=number/count)



df_summary <- df_summary[order(df_summary$year_release),] %>% unique(.) %>% select(year_release,
subject_sex,subject_race,number,proportion)
df_summary
```

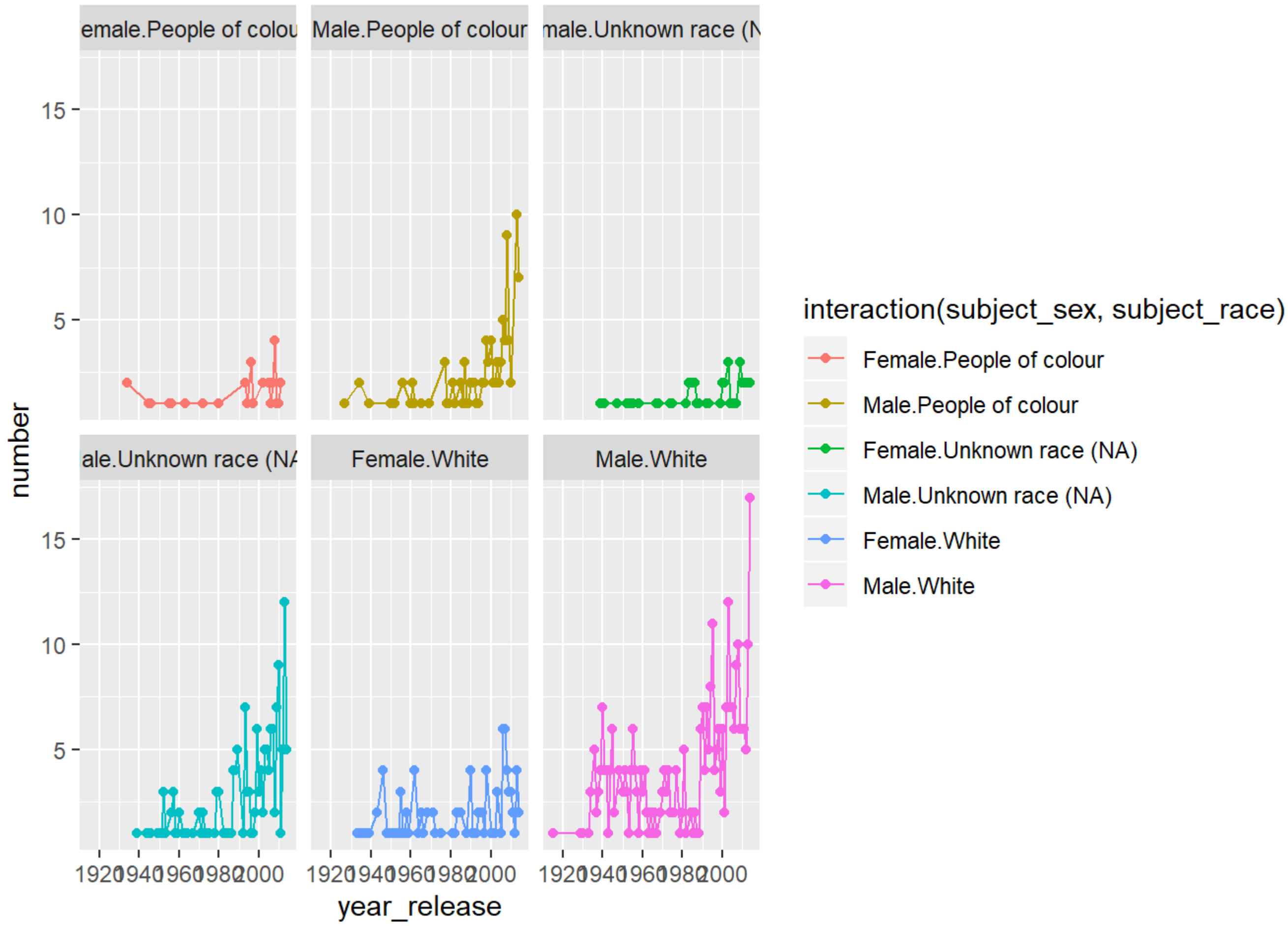
```
## # A tibble: 281 x 5
## # Groups:   year_release, subject_sex, subject_race [281]
##    year_release subject_sex subject_race     number proportion
##           <int> <chr>       <chr>             <int>      <dbl>
##  1         1915 Male        White                 1      1
##  2         1927 Male        People of colour      1      1
##  3         1929 Male        White                 1      1
##  4         1930 Male        White                 1      1
##  5         1933 Female      White                 1      0.5
##  6         1933 Male        White                 1      0.5
##  7         1934 Female      People of colour      2      0.286
##  8         1934 Male        White                 3      0.429
##  9         1934 Male        People of colour      2      0.286
## 10         1935 Female      White                 1      1
## # ... with 271 more rows
```

f. Create (i) a line plot showing the counts of these groups over time and (ii) a line plot showing the relative proportions of subjects over time. Would you infer from these plots that the imbalance is improving over time or not? Explain your answer.

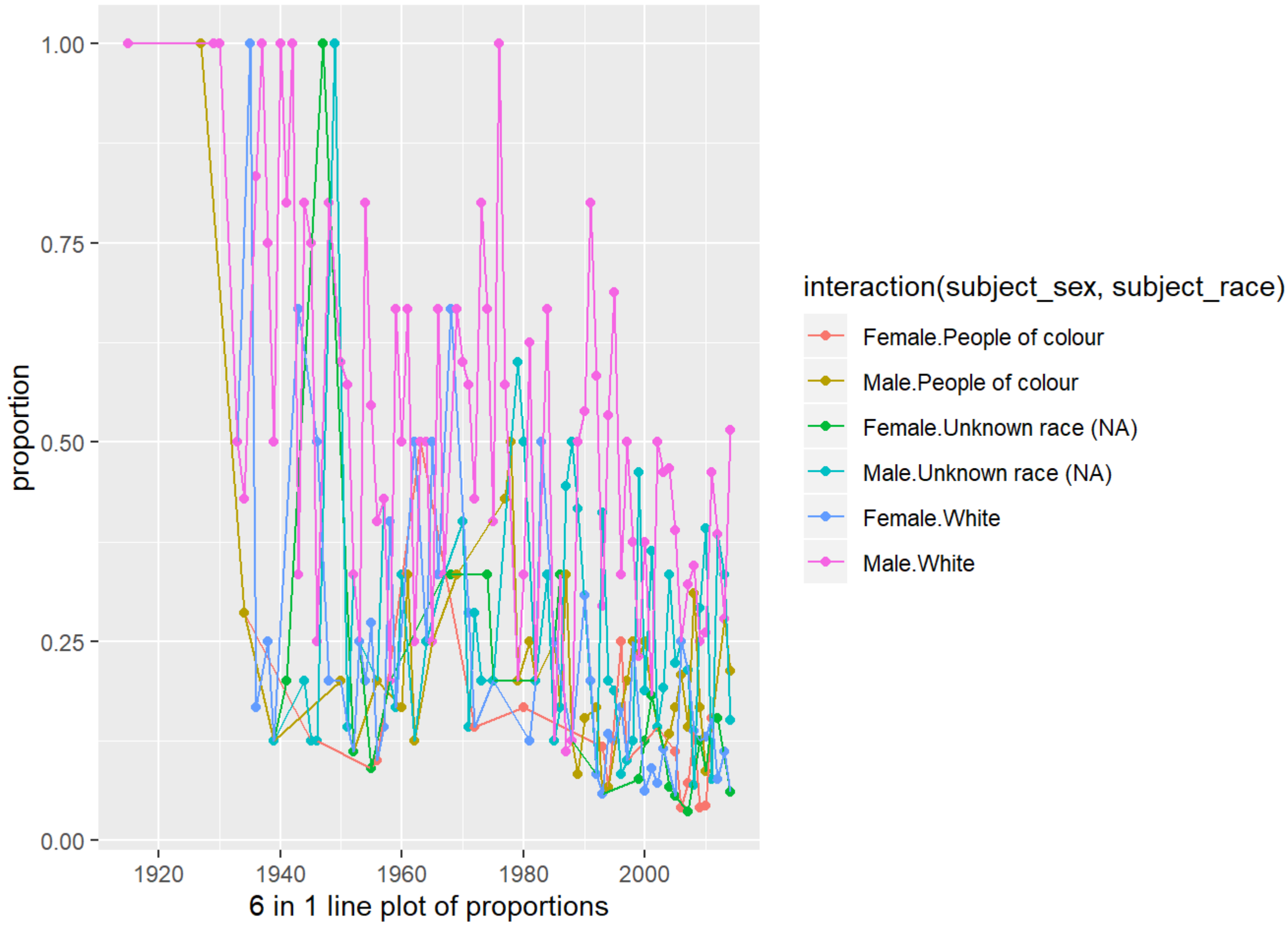
```
#df_summary

ggplot(df_summary,aes(x=year_release,y=number,fill=interaction(subject_sex,subject_race))) +
  geom_point(aes(color=interaction(subject_sex,subject_race))) + geom_line(stat="summary",fun.y=
mean,aes(color=interaction(subject_sex,subject_race))) +
  xlab("6 in 1 line plot of counts")
```



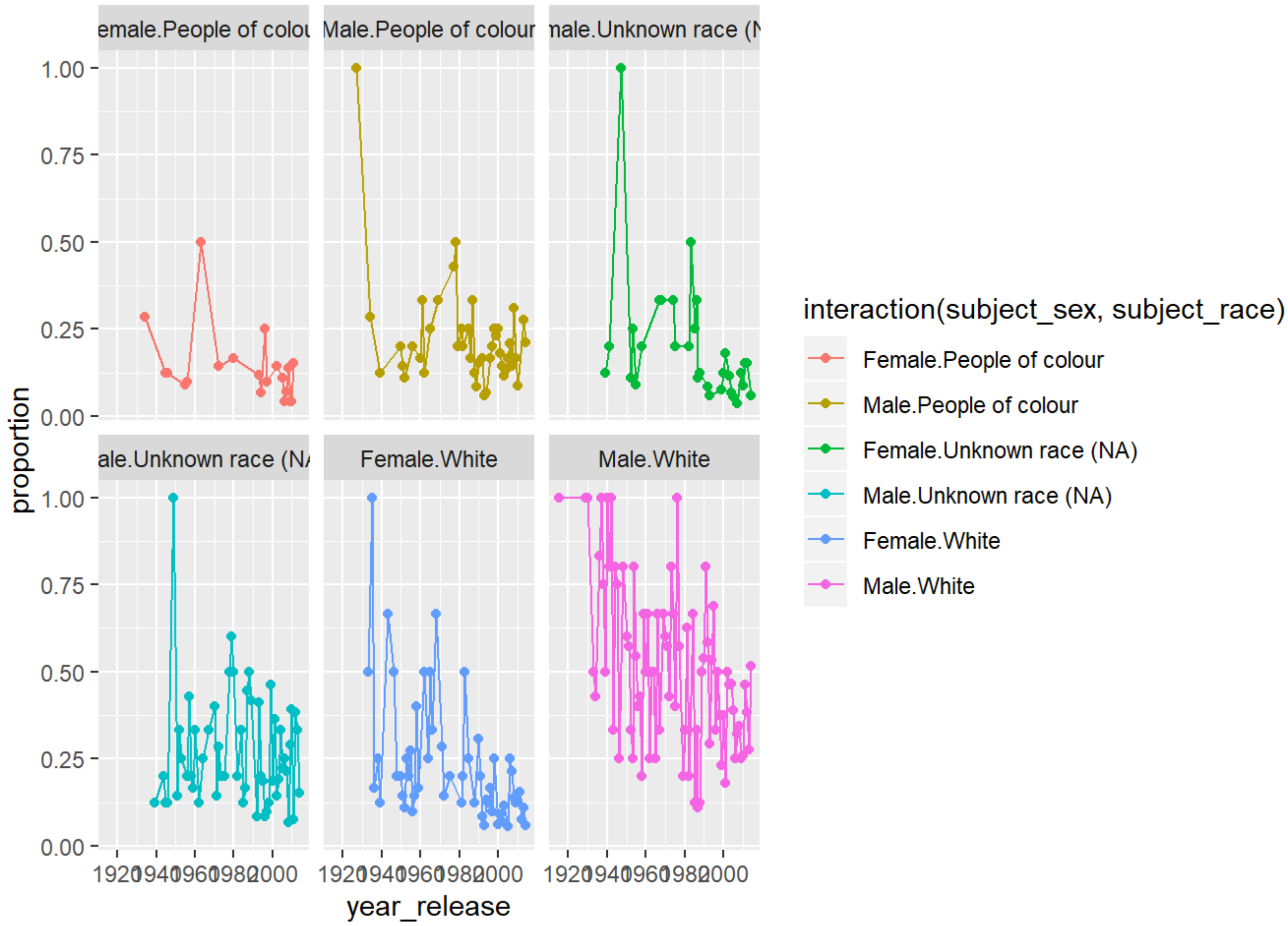

```
ggplot(df_summary,aes(x=year_release,y=number,fill=interaction(subject_sex,subject_race))) +
  geom_point(aes(color=interaction(subject_sex,subject_race))) + geom_line(stat="summary",fun.y=
mean,aes(color=interaction(subject_sex,subject_race))) + facet_wrap(~interaction(subject_sex,sub
ject_race))
```




```
ggplot(df_summary,aes(x=year_release,y=proportion,fill=interaction(subject_sex,subject_race))) +
  geom_point(aes(color=interaction(subject_sex,subject_race))) + geom_line(stat="summary",fun.y=
mean,aes(color=interaction(subject_sex,subject_race))) +
  xlab("6 in 1 line plot of proportions")
```




```
ggplot(df_summary,aes(x=year_release,y=proportion,fill=interaction(subject_sex,subject_race))) +
  geom_point(aes(color=interaction(subject_sex,subject_race))) + geom_line(stat="summary",fun.y=
mean,aes(color=interaction(subject_sex,subject_race))) + facet_wrap(~interaction(subject_sex,sub
ject_race))
```




```
# The problem of imbalance is getting better as we can tell from the line plot of the proportion
s that the percentages are all gradually converging to 1/6.
```

Question 2 (a) First, create a summary table that finds the mean and median for each of the six quantitive variables with a column for each group. (Hint: use summarise, pivot_longer, and pivot_wider). Which varible(s) seem to differentiate amongst the different types of diabetes?

```
library(heplots)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

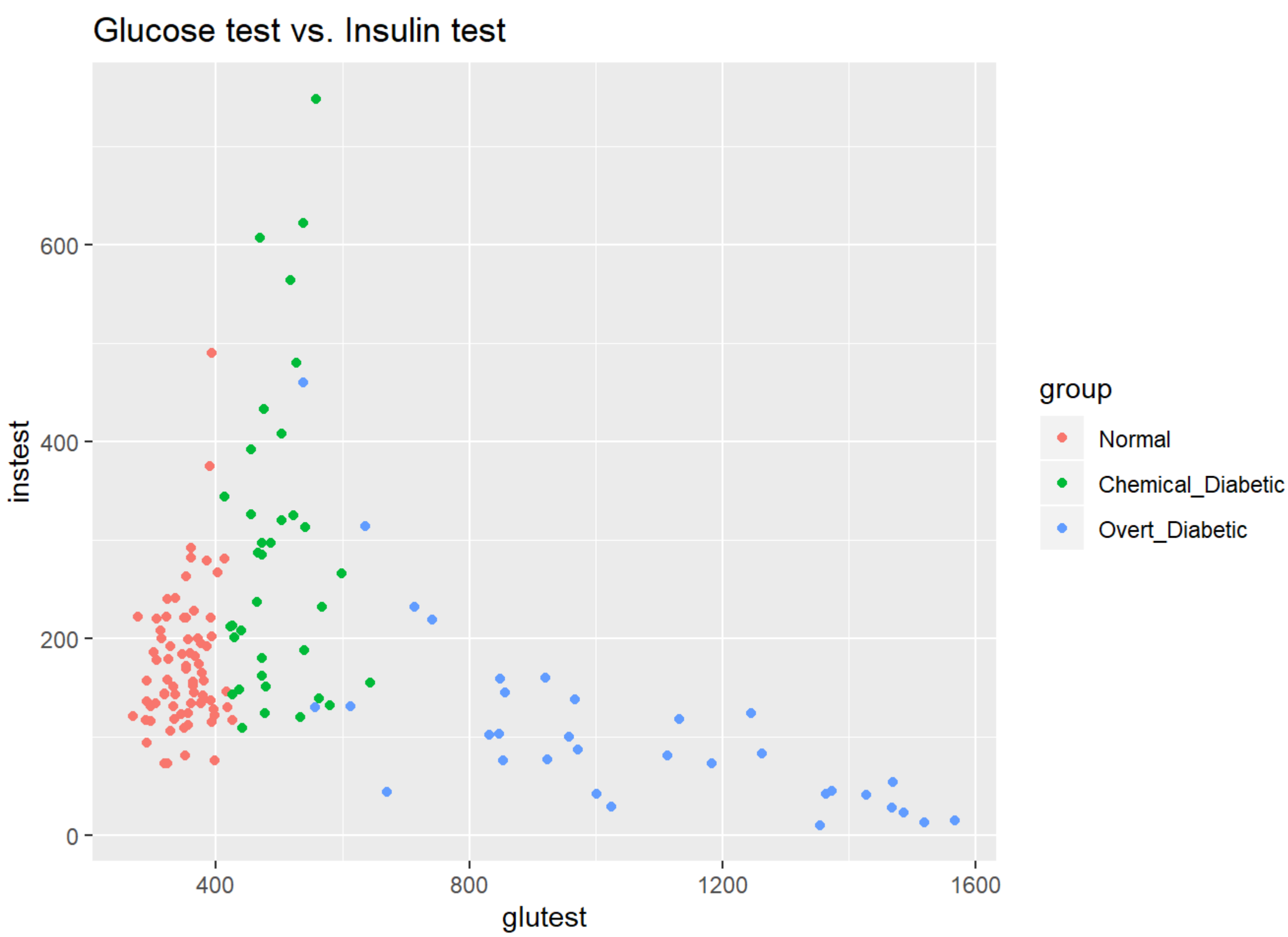
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

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

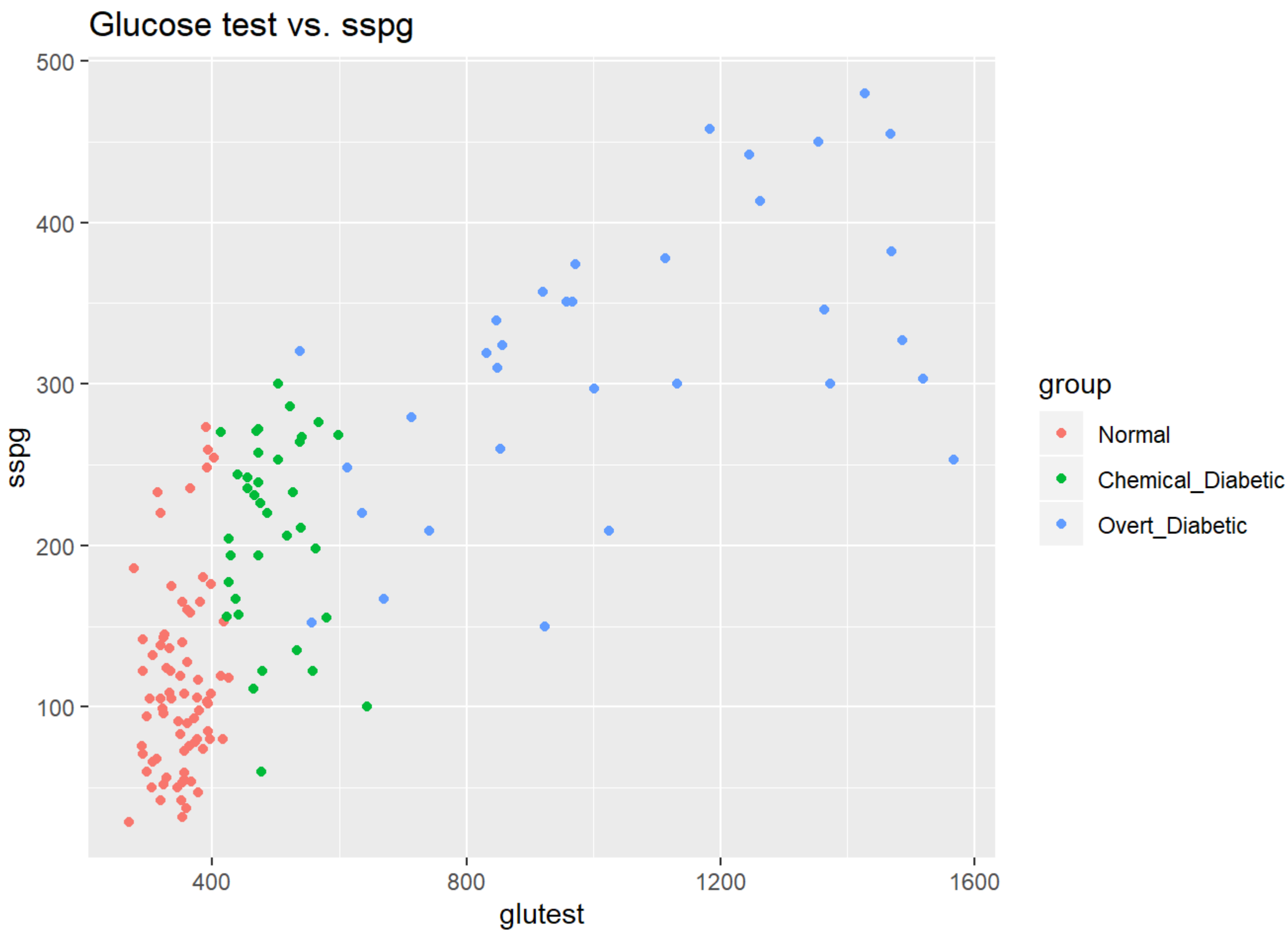
```
df <- tbl_df(Diabetes)
df_grouped <- df %>% group_by(group) %>% summarise_all(list(Avg=mean,Med=median)) %>%
  pivot_longer(cols=contains('_'),names_to = "Measure") %>%
  pivot_wider(id_cols = Measure, names_from = group) %>% arrange(desc(Measure))

# variables that seem to differentiate amongst the different types of diabetes: sspg, instest, g
lutest, glufast.
```

b. Create 3 scatterplots, comparing all possible pairs of the glucose test variable, the insulin test varible and the sspg variable. Which pair of variables seems to allow for the strongest distinction amongst the three groups?

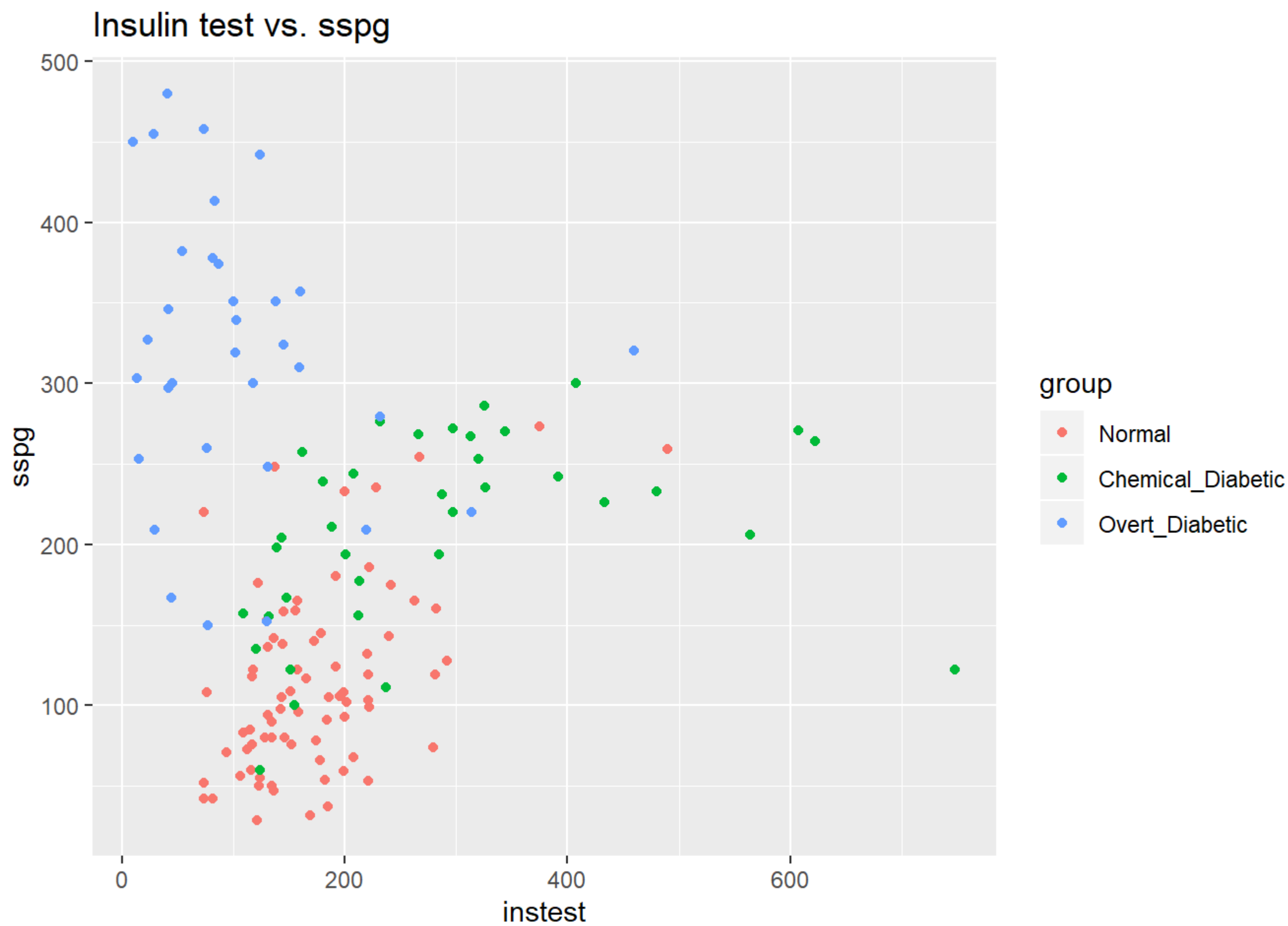
```
#glucose, insulin, sspg


library(heplots)
df <- tbl_df(Diabetes)
df_grouped <- df %>% group_by(group)
ggplot(df_grouped,aes(x=glutest,y=instest,col=group)) + geom_point() +
  labs(x="glutest",y="instest",title="Glucose test vs. Insulin test")
```



```
ggplot(df_grouped,aes(x=glutest,y=sspg,col=group)) + geom_point() +
  labs(x="glutest",y="sspg",title="Glucose test vs. sspg")
```



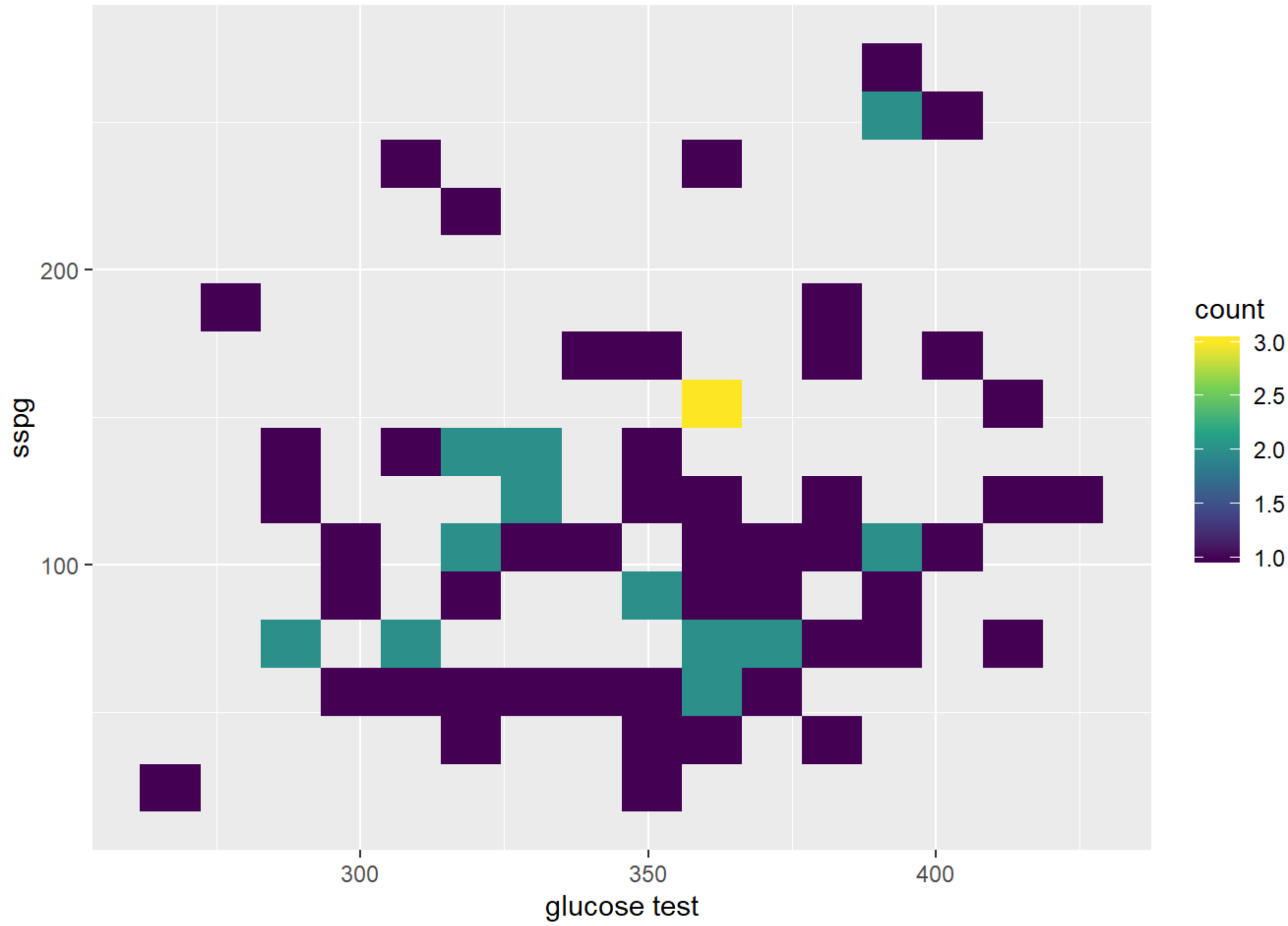

```
ggplot(df_grouped,aes(x=instest,y=sspg,col=group)) + geom_point() +
  labs(x="instest",y="sspg",title="Insulin test vs. sspg")
```



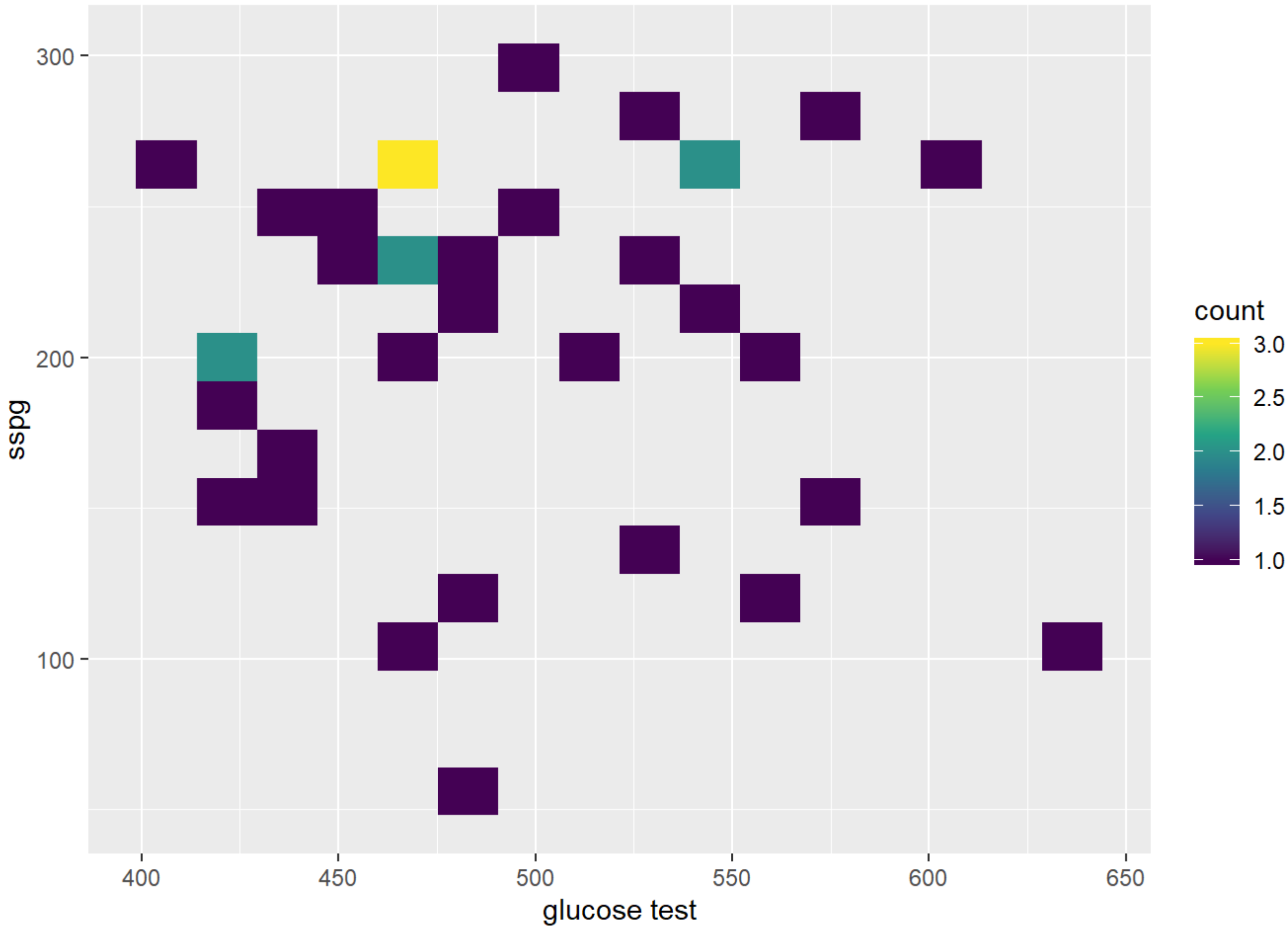

```
# Glucose test & sspg seem to allow for the max distinction because the points are more clustere
d in this case.
```

c. Using the pair of variables that you chose in part (b), make 2-d histograms and contour plots for each group separately. Do you find for this dataset that these plot provide useful summaries of the differences in distributions in the three groups? Feel free to adjust the amount of binning/smoothing and the number of levels from the defalut levels.

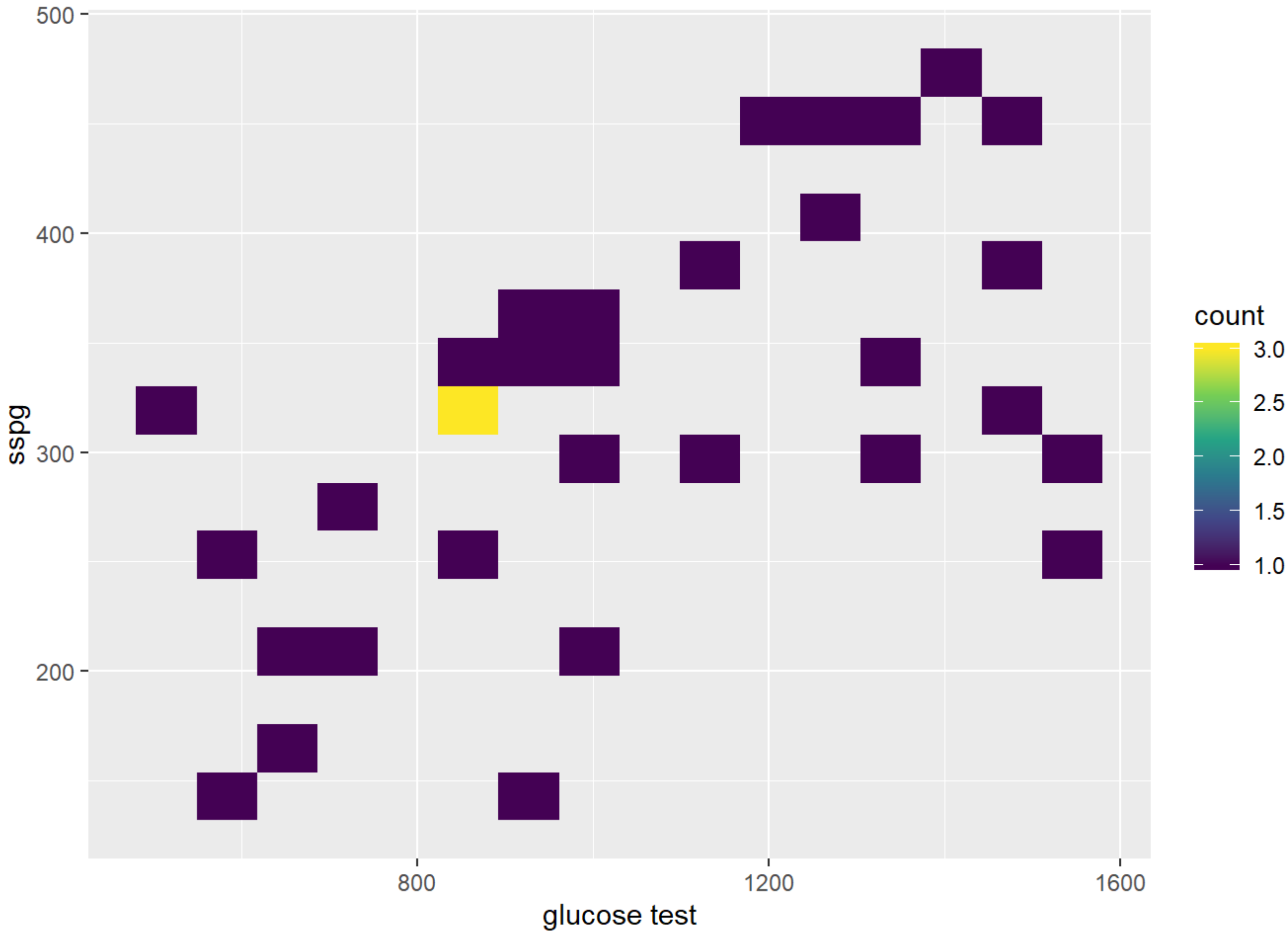
```
df_grouped_normal <- df_grouped[df_grouped$group == "Normal",]
df_grouped_chemical <- df_grouped[df_grouped$group == "Chemical_Diabetic",]
df_grouped_overt <- df_grouped[df_grouped$group == "Overt_Diabetic",]

ggplot(df_grouped_normal,aes(x=glutest,y=sspg)) + geom_bin2d(bins=15)+
  scale_fill_continuous(type="viridis") +
  labs(x="glucose test",y="sspg")
```



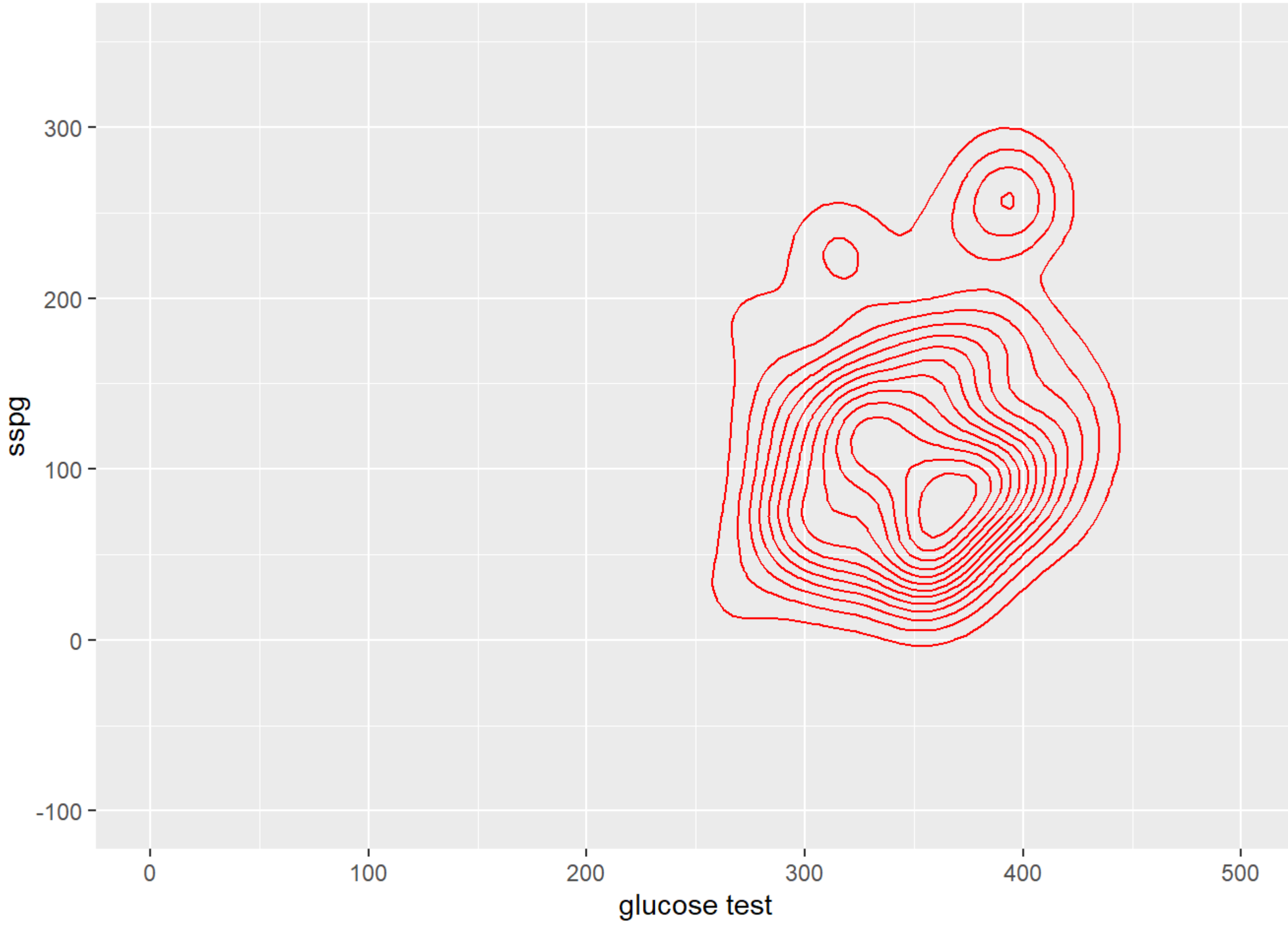

```
ggplot(df_grouped_chemical,aes(x=glutest,y=sspg)) + geom_bin2d(bins=15)+
  scale_fill_continuous(type="viridis") +
  labs(x="glucose test",y="sspg")
```



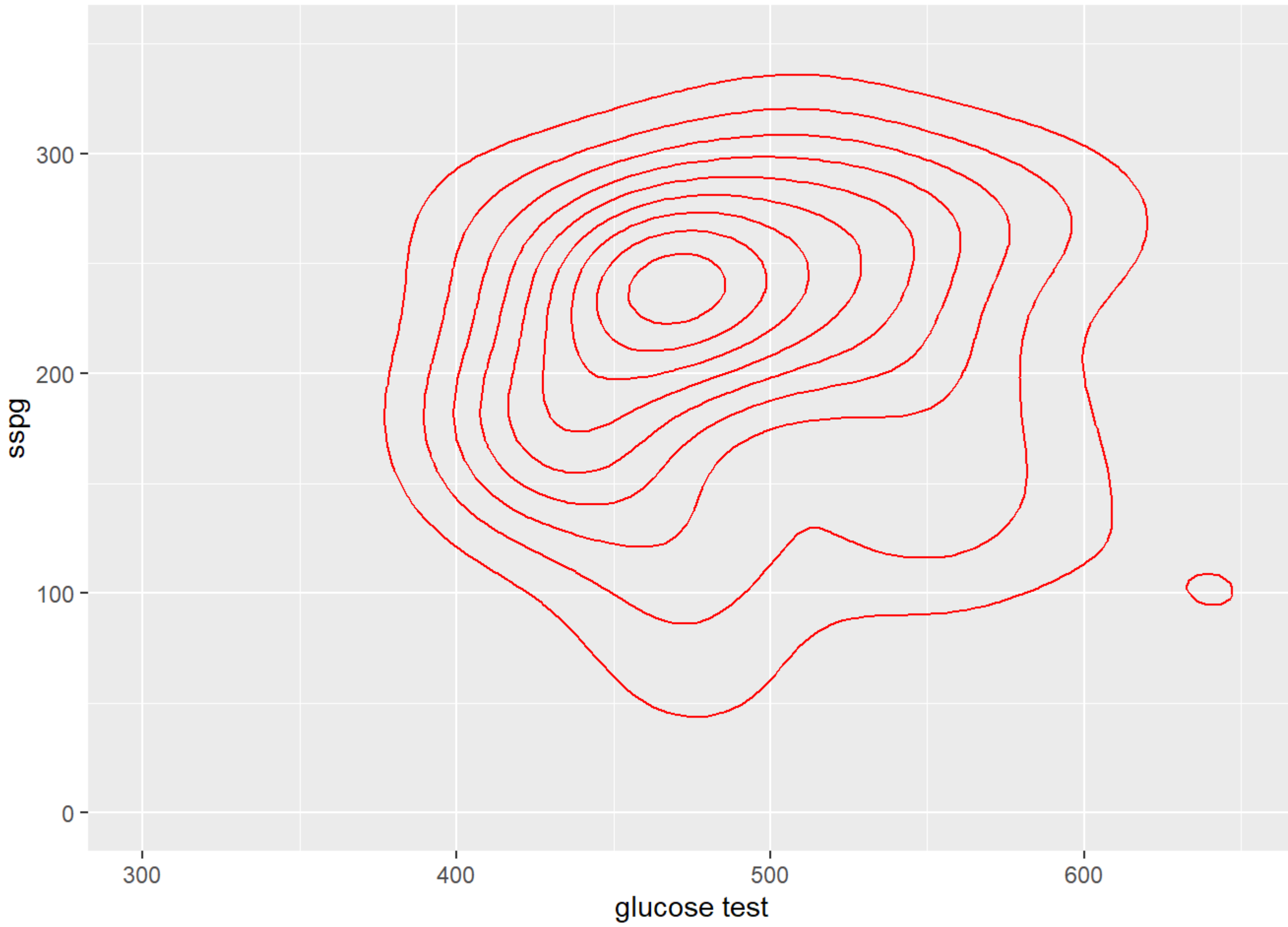

```
ggplot(df_grouped_overt,aes(x=glutest,y=sspg)) + geom_bin2d(bins=15)+
  scale_fill_continuous(type="viridis") +
  labs(x="glucose test",y="sspg")
```



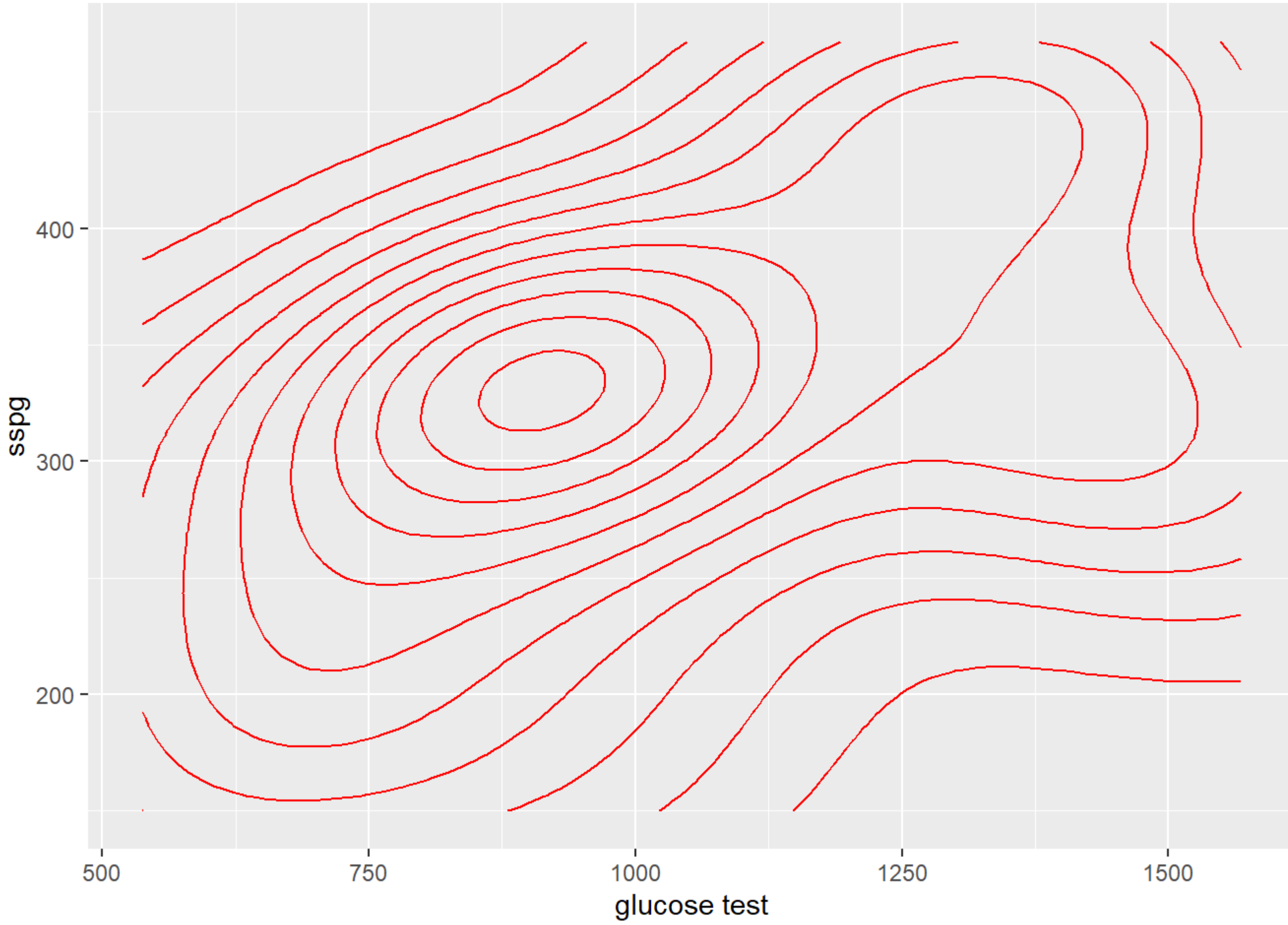

```
ggplot(df_grouped_normal,aes(x=glutest,y=sspg)) + geom_density_2d(col="red") +
  labs(x="glucose test",y="sspg") + ylim(c(-100,350)) + xlim(c(0,500))
```




```
ggplot(df_grouped_chemical,aes(x=glutest,y=sspg)) + geom_density_2d(col="red") +
  labs(x="glucose test",y="sspg") + ylim(c(0,350)) + xlim(c(300,650))
```




```
ggplot(df_grouped_overt,aes(x=glutest,y=sspg)) + geom_density_2d(col="red") +
  labs(x="glucose test",y="sspg") #+ ylim(c(400,1000)) + xlim(c(0,500))
```




```
print(df_grouped_overt)
```

```
## # A tibble: 33 x 6
## # Groups:   group [1]
##    relwt glufast glutest instest  sspg group
##    <dbl>   <int>   <int>   <int> <int> <fct>
##  1  0.92     300    1468      28   455 Overt_Diabetic
##  2  0.86     303    1487      23   327 Overt_Diabetic
##  3  0.85     125     714     232   279 Overt_Diabetic
##  4  0.83     280    1470      54   382 Overt_Diabetic
##  5  0.85     216    1113      81   378 Overt_Diabetic
##  6  1.06     190     972      87   374 Overt_Diabetic
##  7  1.06     151     854      76   260 Overt_Diabetic
##  8  0.92     303    1364      42   346 Overt_Diabetic
##  9  1.2      173     832     102   319 Overt_Diabetic
## 10  1.04     203     967     138   351 Overt_Diabetic
## # ... with 23 more rows
```