Homework #2

Chi Zhang

Chapter 4

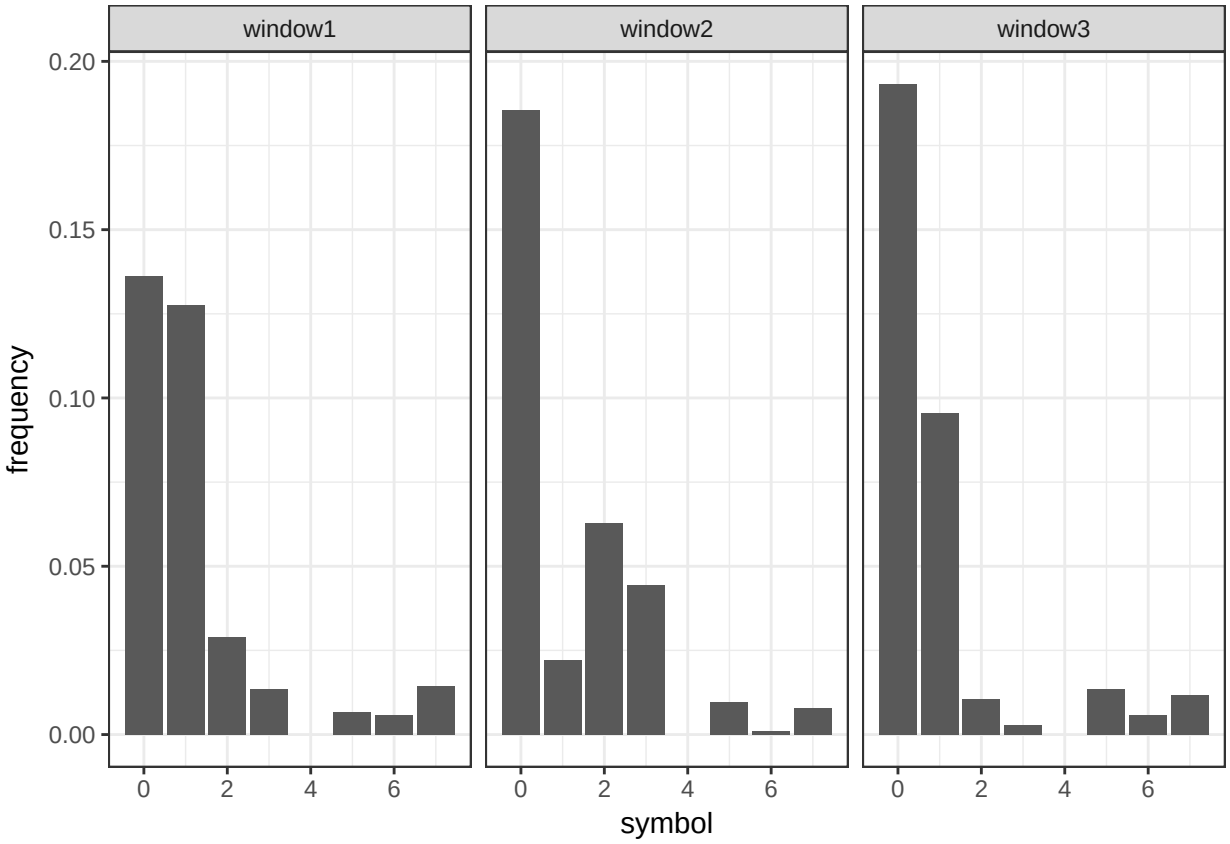
1. Slot Machines (Chapter 4 exercises, #3, p. 72)

[5 points]

Do not use *grid.arrange()* for this exercise. Rather, use *gather()* to tidy the data and then facet on window number. To make the comparison, use relative frequency bar charts (the heights of the bars in each facet sum to one). Describe how the distributions differ.

```
library('DAAG')
library('tidyr')
library('ggplot2')
mydata1<-vlt[c('window1','window2','window3')] %>%
gather(window,value) %>%
ggplot(aes(value,window))+geom_bar(aes(y=(..count..)/sum(..count..)))+facet_wrap(~window)+
  ylab('frequency')+xlab('symbol')+theme_bw()

mydata1
```



Relative frequency of symbol 0 is quite different between window1 and window3. It has highest frequency in window3 and lowest in window1. And for symbol 1, it has second highest relative frequency in window1 and window3.

2. Detailed Mortality data ("Death2015.txt")

[21 points]

This data comes from the "Detailed Mortality" database available on https://wonder.cdc.gov/

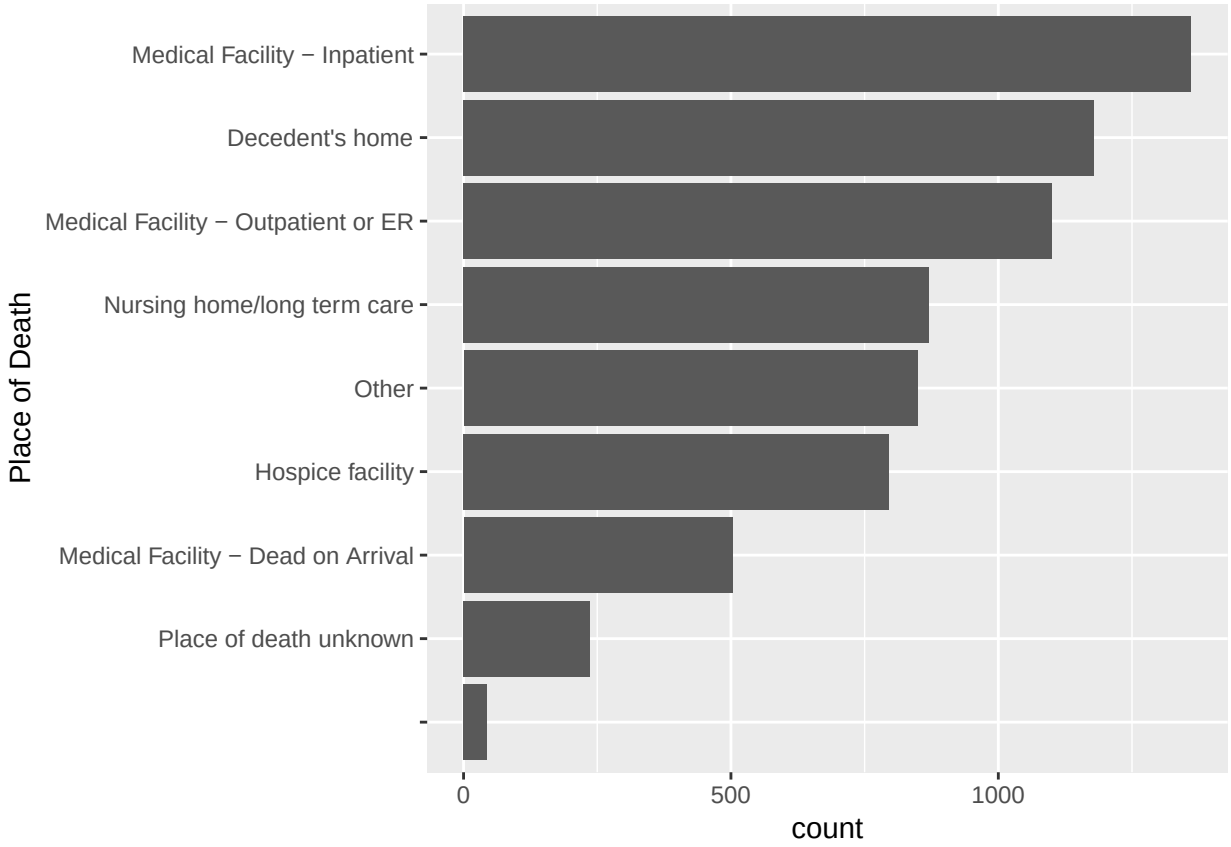
Code for all preprocessing must be shown. (That is, don't open in the file in Excel or similar, change things around, save it, and then import to R. Why? Because your steps are not reproducible.)

(a) For `Place of Death`, `Ten-Year Age Groups`, and `ICD Chapter Code` variables, do the following:

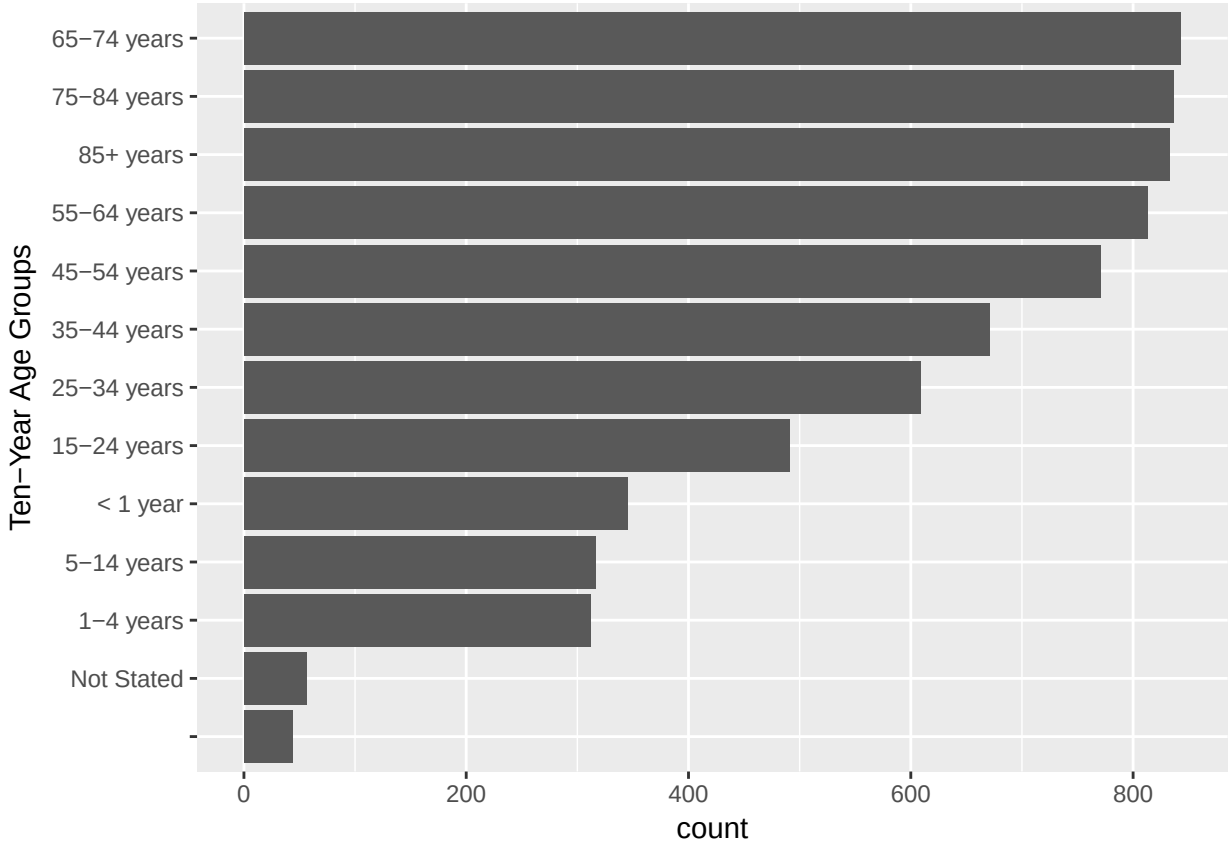
Identify the type of variable (nominal, ordinal, or discrete) and draw a horizontal bar chart using best practices for order of categories.

Place of death: nominal Ten-year age groups: ordinal ICD chapter code: nominal

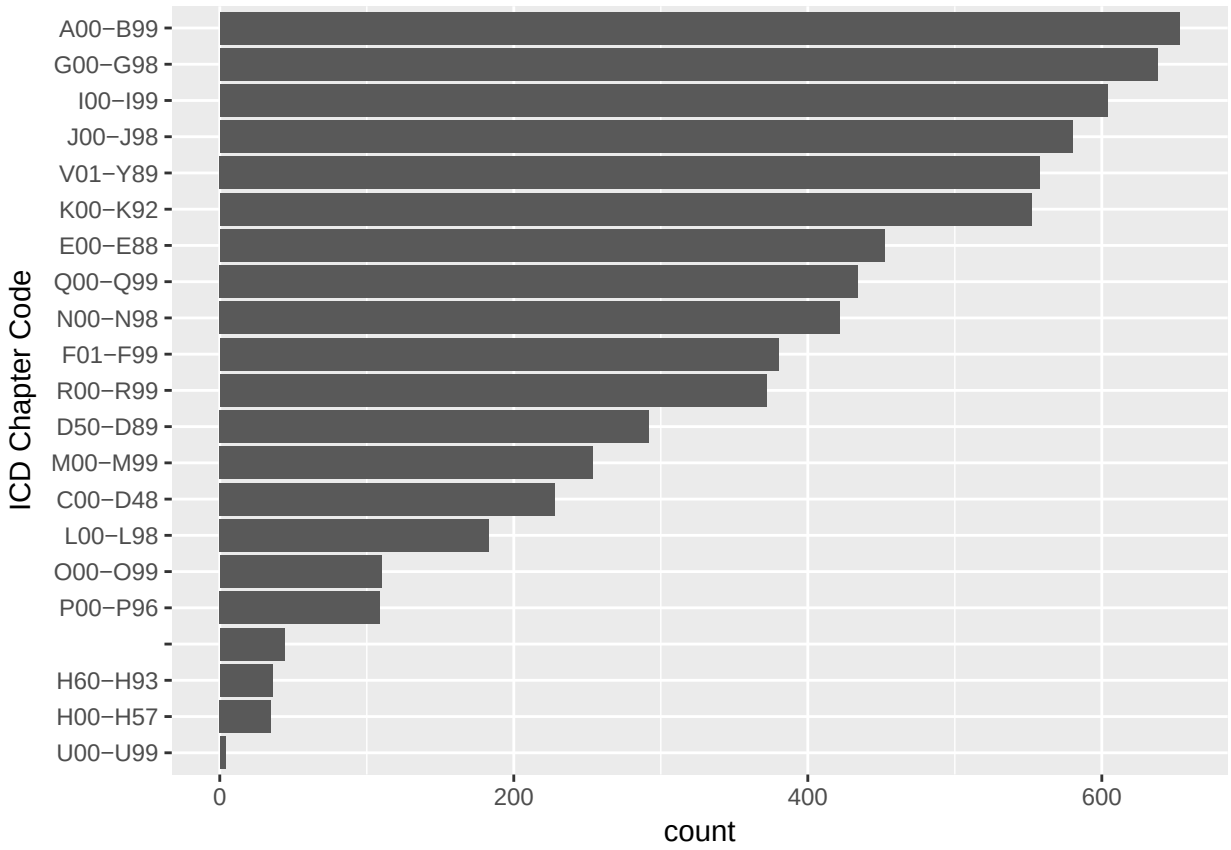
```
library('dplyr')
mydata2 <- read.delim("/Users/albertzhang/Desktop/18spring/EDAV/HW/HW2/Death2015.txt")
group1 <- mydata2 %>% group_by(Place.of.Death) %>% summarize(count = n())
ggplot(group1, aes(x = reorder(Place.of.Death, count), y = count)) + geom_col() + xlab("Place of Death")
```




```
group2 <- mydata2 %>% group_by(Ten.Year.Age.Groups) %>% summarize(count = n())
ggplot(group2, aes(x = reorder(Ten.Year.Age.Groups, count), y = count)) + geom_col() +
  xlab("Ten-Year Age Groups")+ coord_flip()
```




```
group3 <- mydata2 %>% group_by(ICD.Chapter.Code) %>% summarize(count = n())
ggplot(group3, aes(x = reorder(ICD.Chapter.Code, count), y = count)) + geom_col()  +
  xlab("ICD Chapter Code") + coord_flip()
```



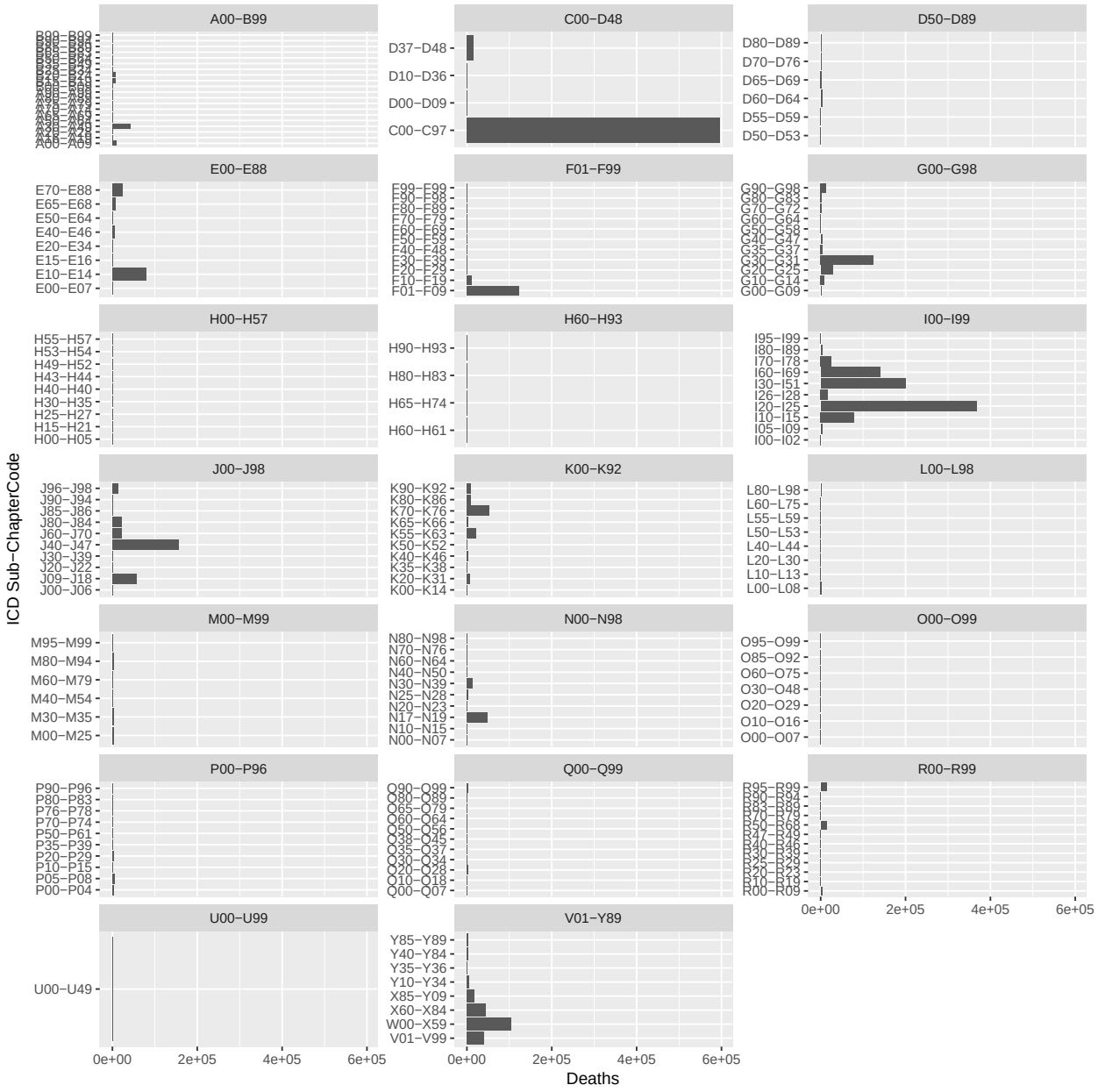
(b) Create horizontal bar charts for the ICD sub-chapter codes, one plot per ICD chapter code, by faceting on chapter code, *not* by using *grid.arrange()*. Use `scales = "free"` with `facet_wrap()`. It should look like this (with data, of course!). Describe notable features.

```
library(dplyr)
mydata2 <- na.omit(mydata2)
mydata2_count <- group_by(mydata2,ICD.Chapter.Code,ICD.Sub.Chapter.Code) %>%
  summarize(count = sum(Deaths))
ggplot(mydata2_count,aes(x=ICD.Sub.Chapter.Code,y=count))+geom_bar(stat = "identity") + coord_flip() +
  facet_wrap(~ICD.Chapter.Code, scales="free",ncol = 3) + theme_grey() + xlab("ICD Sub-Chapter Code") +
```

Most A00-B99 data comes from subchapterA30-A49; Most C00-D48 data comes from subchapterC00-C97; Most F01-F99 data comes from subchapterF01-F09; Nearly all data comes from U00-U49

(c) Change the `scales` parameter to `scales = "free_y"`. What changed? What information does this set of graphs provide that wasn't available in part (b)?

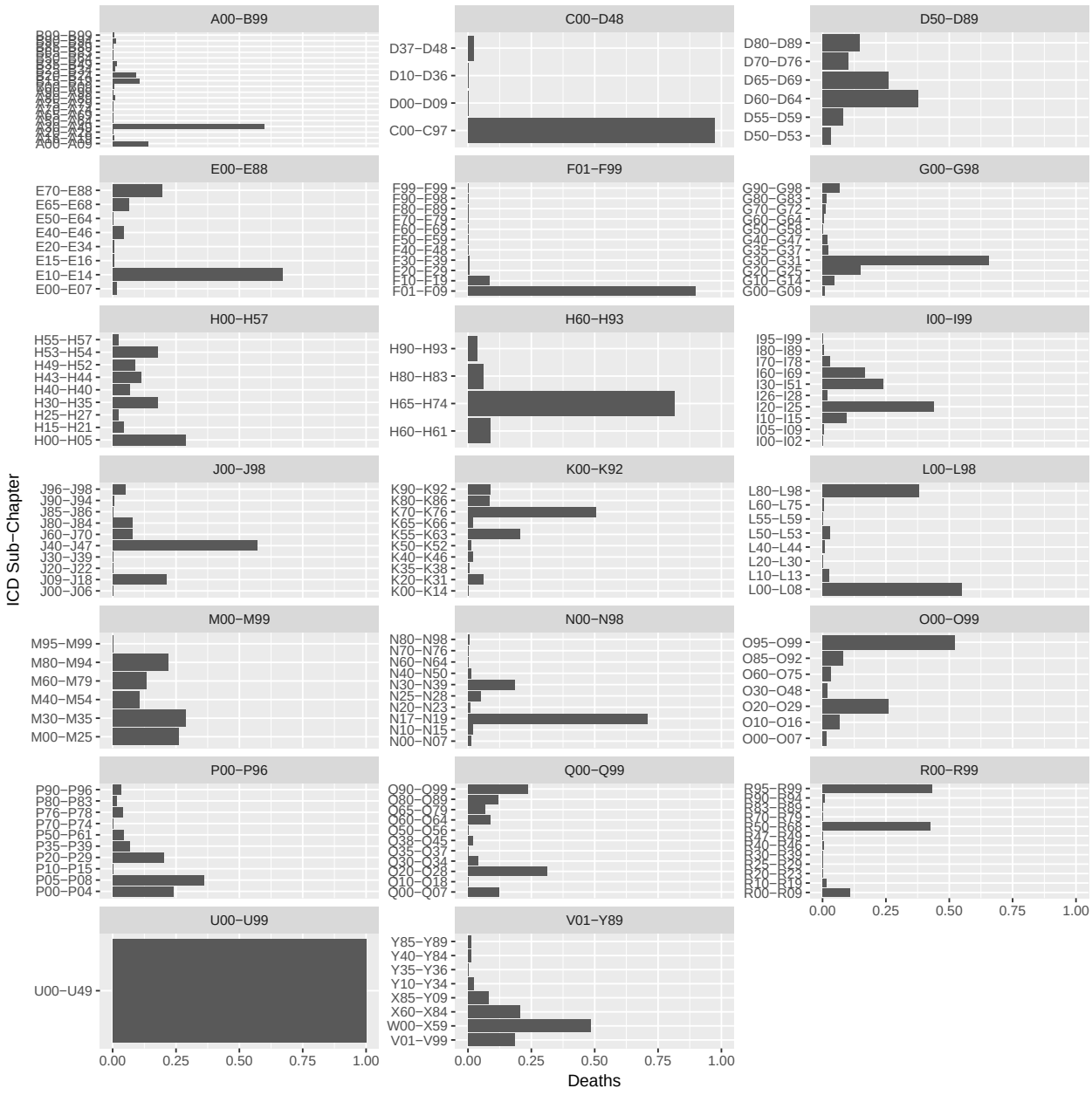
```
library(dplyr)
ggplot(mydata2_count,aes(x=ICD.Sub.Chapter.Code,y=count))+geom_bar(stat = "identity")+
  coord_flip()+facet_wrap(~ICD.Chapter.Code,scales="free_y",ncol = 3)+theme_grey()+xlab("ICD Sub-Chapte
```



Only sub-chapter code is changed to free and this situation changes in terms of different plots. Now we can easily compare count in each panel since the scale is different.

(d) Redraw the panels as *relative frequency* bar charts rather than *count* bar charts. (The lengths of the bars *in each panel separately* must sum to 1.) What new information do you gain?

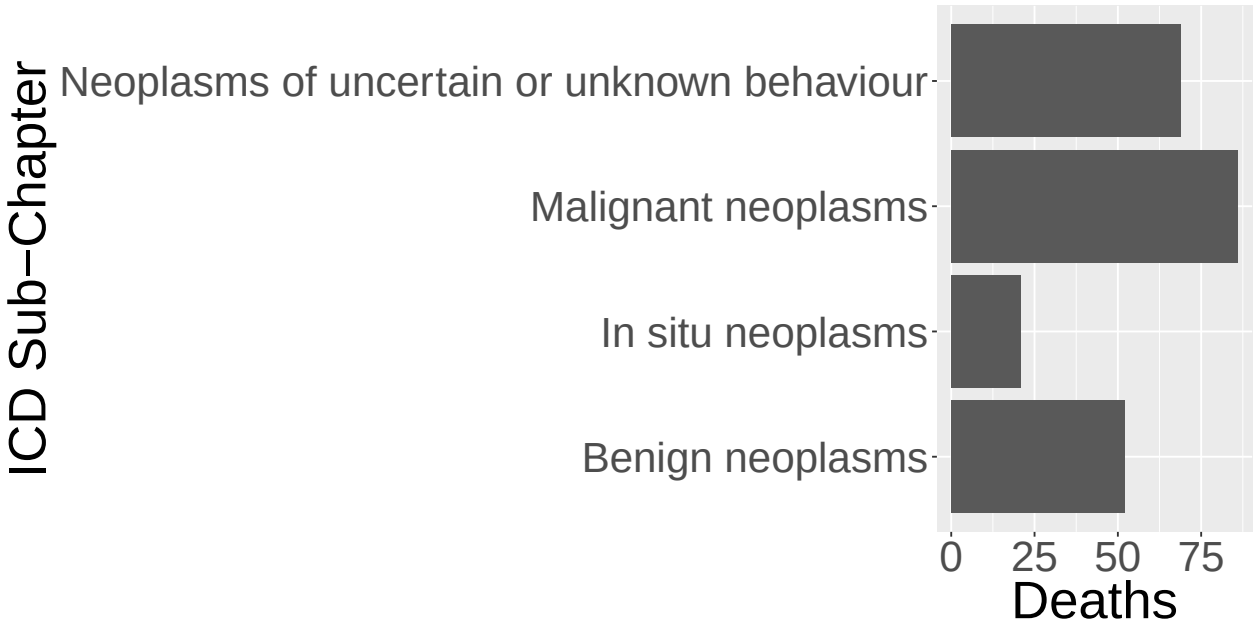
```
library(dplyr)
mydata3<-aggregate(Deaths~ICD.Sub.Chapter.Code+ICD.Chapter.Code,mydata2, sum)
ggplot(mydata3, aes(x =ICD.Sub.Chapter.Code ,
                    y = (Deaths)/sapply(PANEL, FUN=function(x) sum(Deaths[PANEL == x])))) +
  geom_col() + facet_wrap(~ICD.Chapter.Code, scales = 'free_y',ncol = 3) + coord_flip()+
  xlab("ICD Sub-Chapter") + ylab("Deaths")
```



By doing this, we can also compare different individual groups, and also could compare different subchapters differ, and how their distribution different from each other chapters.

(e) Choose one of the small panels and redraw it as a single graph, using names rather than codes. (That is, use `ICD Chapter` and `ICD Sub-Chapter` instead of the code versions.) What type of data is this? Note any interesting features.

```
mydata3 <- data.frame(mydata2$ICD.Sub.Chapter[which(mydata2$ICD.Chapter == 'Neoplasms')])
names(mydata3) = 'Neoplasms'
ggplot(mydata3, aes(Neoplasms)) + geom_bar() + xlab("ICD Sub-Chapter") + ylab("Deaths") +
  theme(text = element_text(size = 30)) + coord_flip()
```



Choose Neoplasms. A nominal feature. Least people die of in situ neoplasms while other three have relatively same mortality.

3. Detailed Mortality, questions about the data

[6 points]

Cite your sources with links.

(a) Who is included in the death counts?

It is based on information from all death certificates filed in the fifty states and the District of Columbia. Deaths of nonresidents (e.g. nonresident aliens, nationals living abroad, residents of Puerto Rico, Guam, the Virgin Islands, and other territories of the U.S.) and fetal deaths are excluded

Citation: Source: Center for Disease Control and Prevention URL: https://wonder.cdc.gov/wonder/help/ucd.html#Mortality.Data

(b) When was this query processed? (Hint: it's in the file itself; don't provide the file time stamp.)

Query Date: Feb 5, 2018 5:08:43 PM.

(c) What does "ICD" stand for? Which version is used for this particular dataset? Name five other countries that use the ICD for official mortality data.

ICD stands for International Classification of Diseases.

Tenth edition version is used.

UK, Canada, China, Korea, and Germany.

Citation: Source: AHIMA URL: http://library.ahima.org/doc?oid=58621#.WoMWt66nG4Q

(d) Which U.S. organizations collects mortality data? Where is the headquarters located?

1. Centers for Disease Control and Prevention.
2. 1600 Clifton Road Atlanta, GA 30329-4027 USA.

Citation: Source: Centers for Disease Control and Prevention URL: https://www.cdc.gov/

(e) In brief, how is the data collected? What is the estimated accuracy rate, according to the dataset documentation?

95% confidence intervals and standard errors for death rates.

Chapter 5

1. Movie ratings

[12 points]

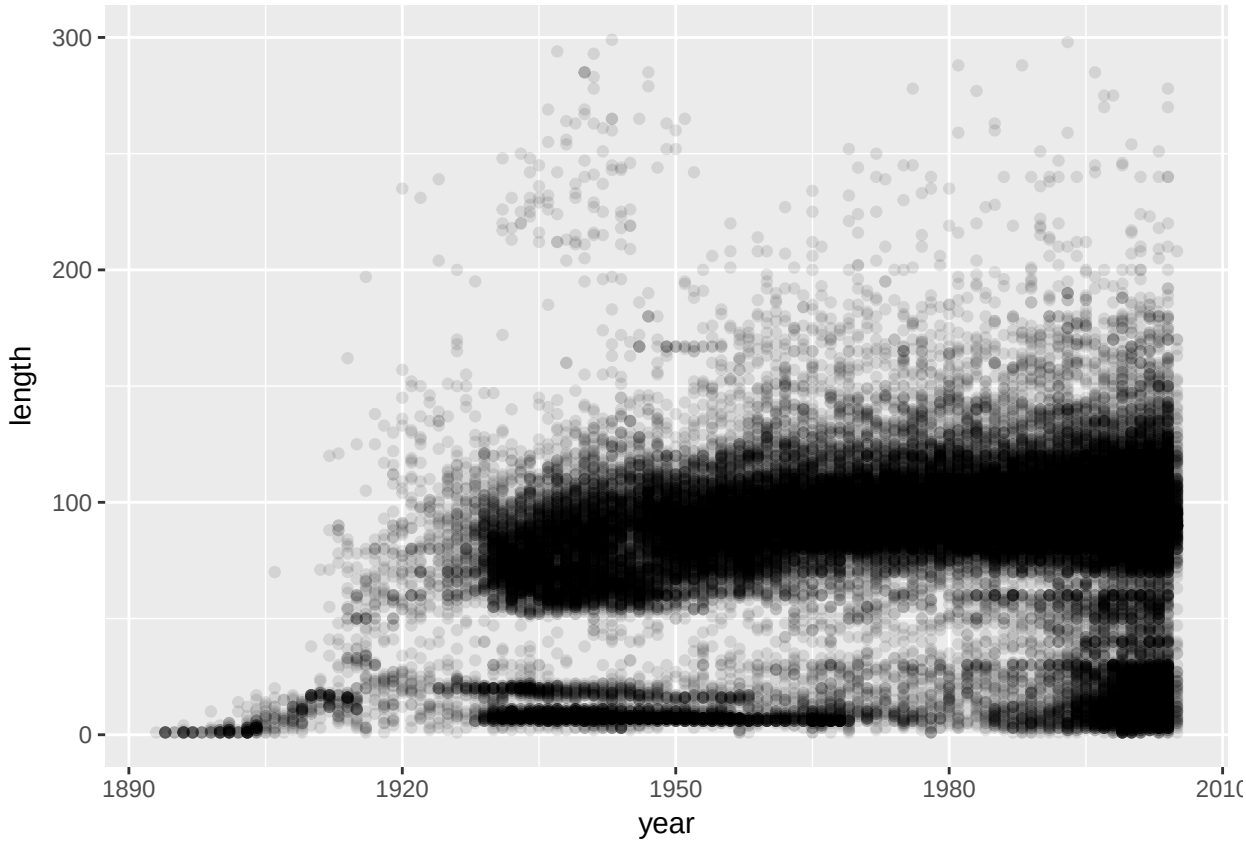
Explore *length* vs. *year* in the **ggplot2movies** dataset, after removing outliers. (Choose a reasonable cutoff).

Draw four scatterplots of *length* vs. *year* from the with the following variations:

(a) Points with alpha blending

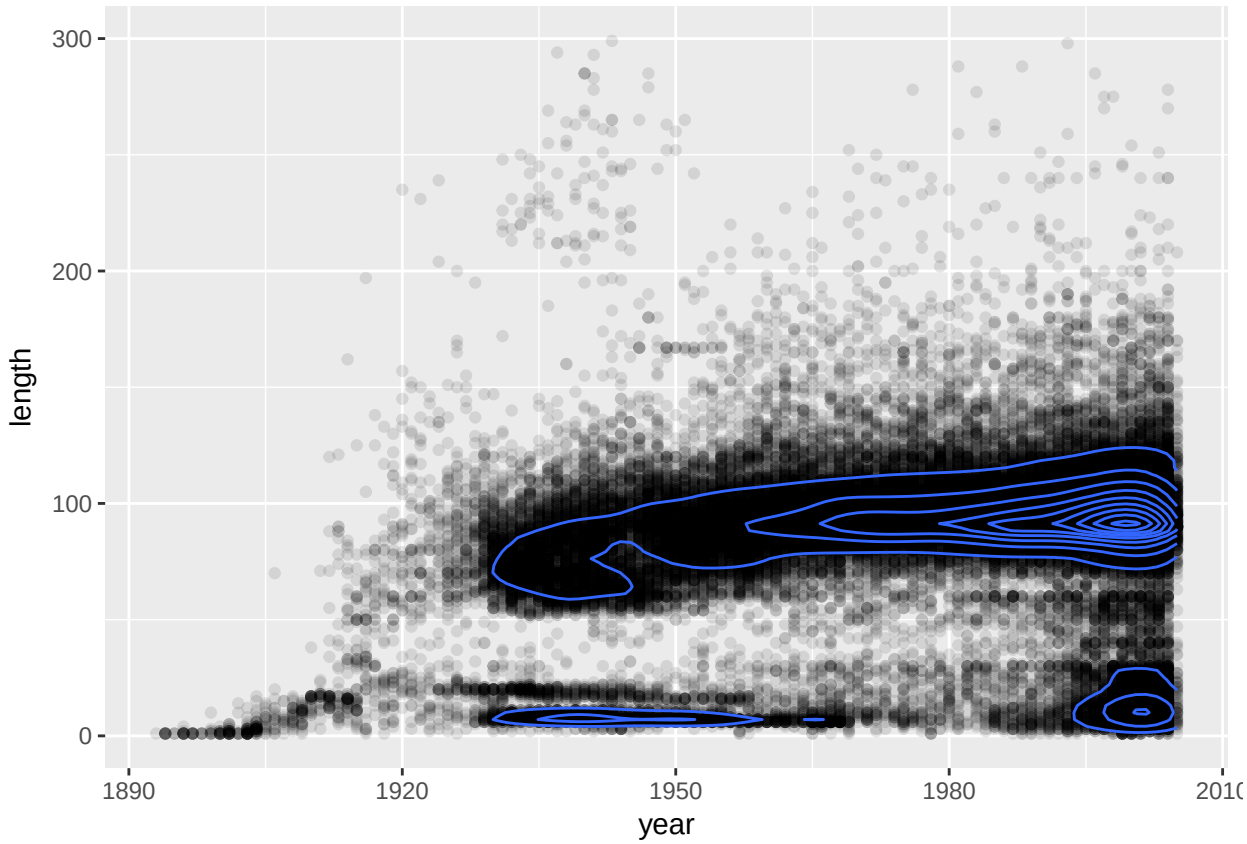
```
library(ggplot2movies)
mydata4 <- movies
data <- subset(mydata4, mydata4$length < 300)[,c("length", "year")]

ggplot(data,aes(year,length))+geom_point(alpha=.1)+theme_grey()
```



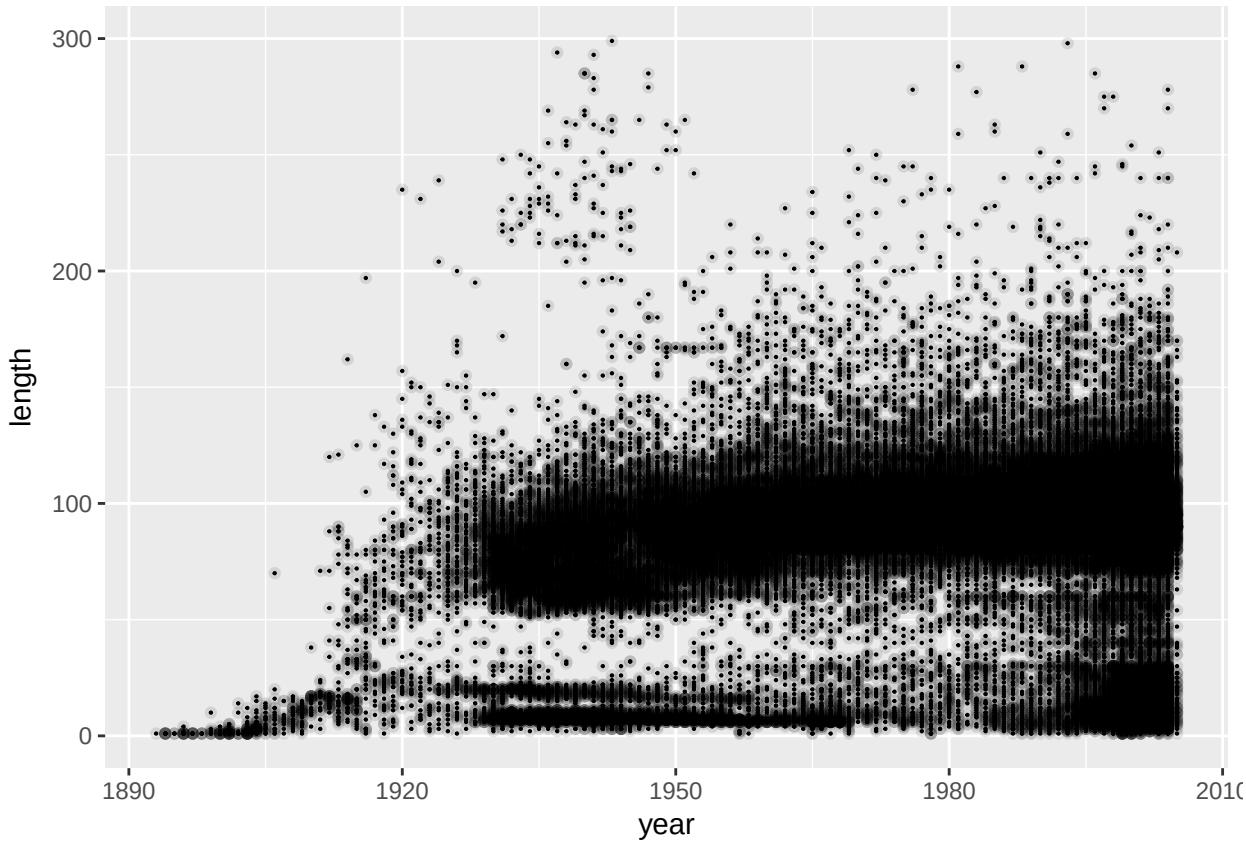
(b) Points with alpha blending + density estimate contour lines

```
ggplot(data,aes(year,length))+geom_point(alpha=.1)+
  theme_grey()+geom_density_2d()
```



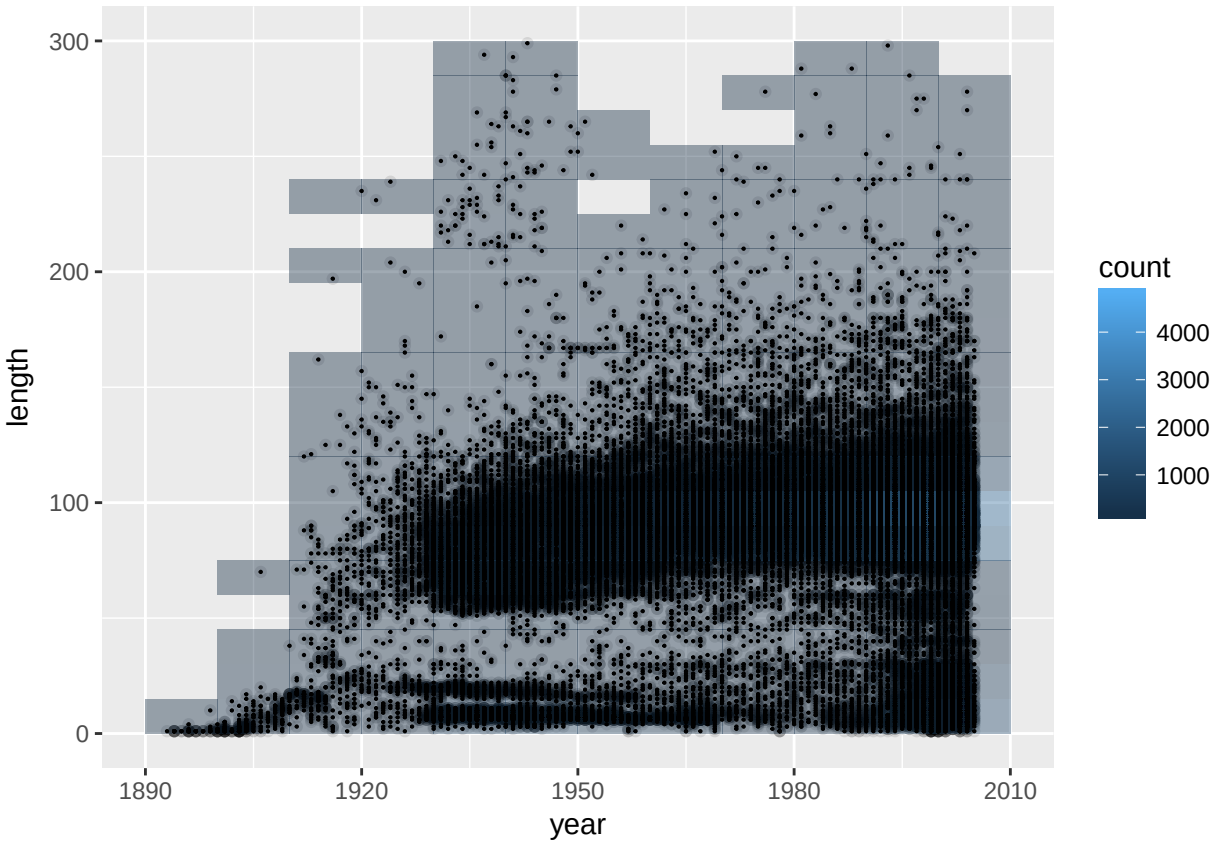
(c) Hexagonal heatmap of bin counts

```
ggplot(data,aes(year,length))+geom_point(alpha=.1)+
  theme_grey()+geom_hex(binwidth = c(10, 15), alpha = 0.5) +
  geom_point(size = 0.1)
```



(d) Square heatmap of bin counts

```
ggplot(data,aes(year,length))+geom_point(alpha=.1)+
  theme_grey() + geom_bin2d(binwidth = c(10,15), alpha = .4) +
  geom_point(size = 0.1)
```



For all, adjust parameters to the levels that provide the best views of the data.

(e) Describe noteworthy features of the data, using the movie ratings example on page 82 (last page of Section 5.3) as a guide.

1.Early movies often have a quite short length. However, the length starts to increase since 1900 and be stable around 100.

2. Throughout all years, least movies have length around 40.

3. Besides 100, there also is another relatively small common length lies around 20.

(f) How do (a)-(d) compare? Are there features that you can see in some but not all of the graphs?

1. Due to large size of dataset, only adjusting alpha value is still not enough to see the most concentrated region. But compared to (a), all (b)(c)(d) give us a better vision on the most concentrated region.

2. (b) could also show regions that have same concentration.

3. (c)(d) have same purpose but (c) could show more clear boundary.

2. Leaves (Chapter 5 exercises, #7, p. 96)

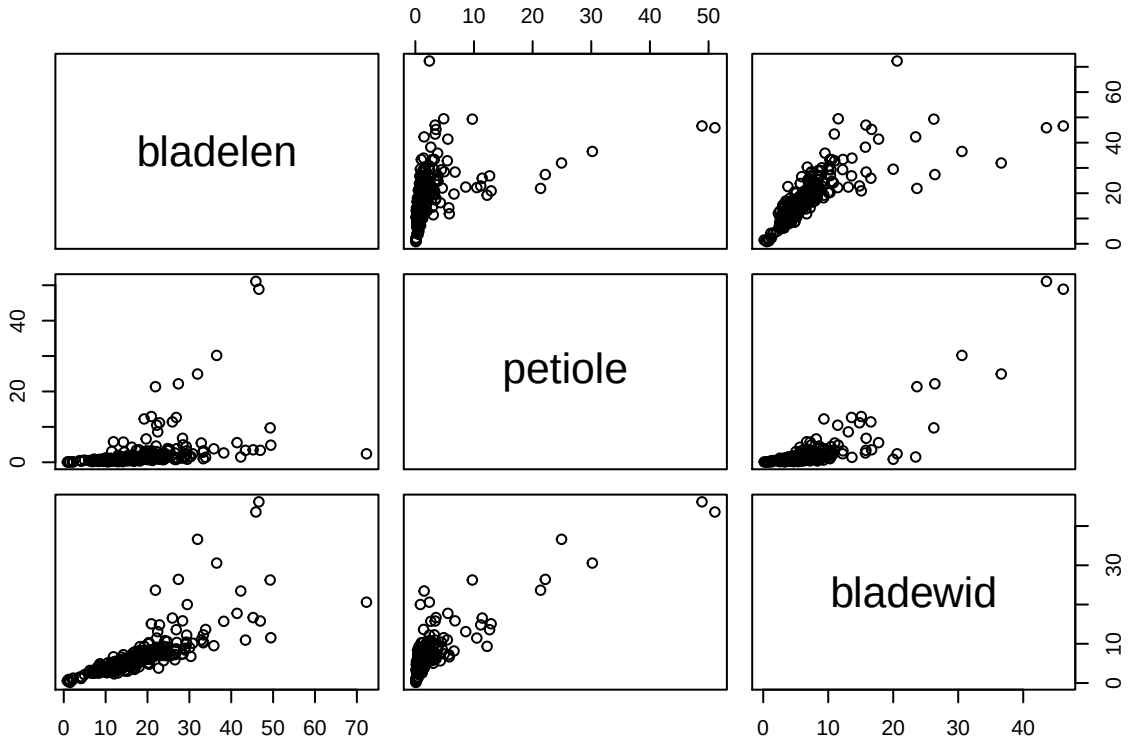
[6 points]

(a)

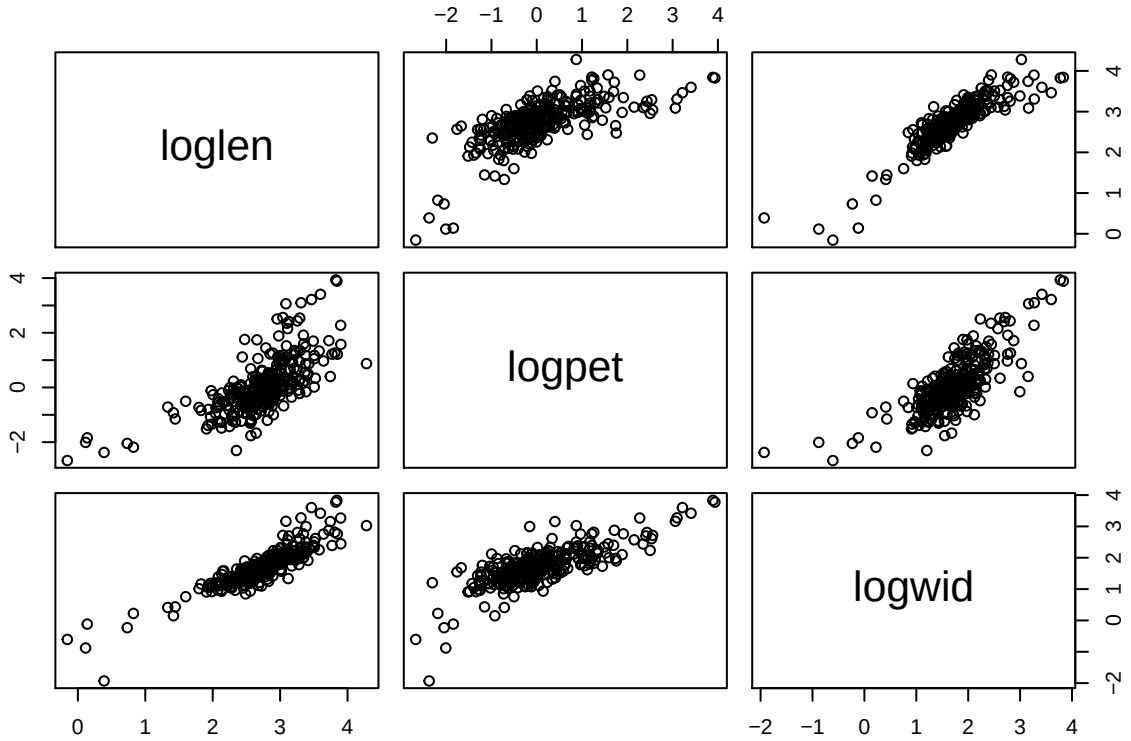
```
library(GGally)
data <- leafshape

splomvar <- data %>% dplyr::select(bladelen, petiole, bladewid)
```

```
splomlogvar <- data %>% dplyr::select(loglen, logpet, logwid)
plot(splomvar)
```




```
plot(splomlogvar)
```



From the scatterplot, we observe that there is a positive correlation between length, width, petiole of leaves.

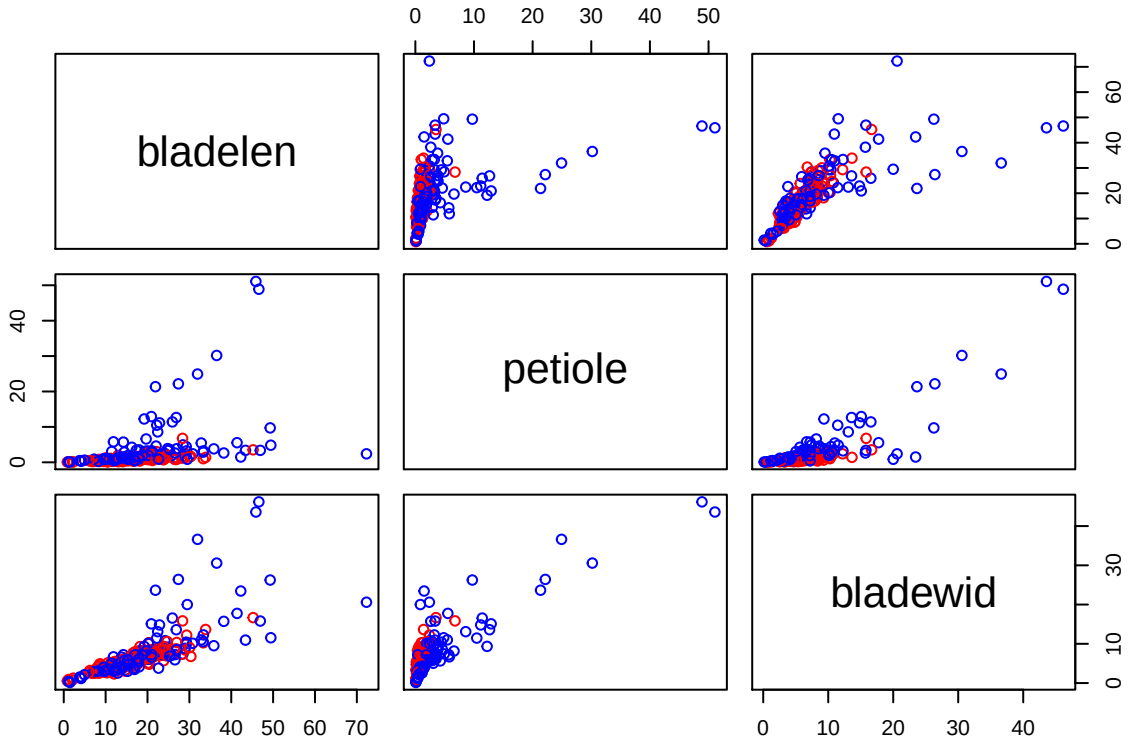
Logarithms of the three measurements are more useful. From the plot without log transformation. We can observe that most of the data are disbributed at the left bottom corner of the plot and the linear tredn is not very good.

After, using the log data. The data are more spread out and the scale of the plot is better. The relationship is more linear and the correlation is more obvious. The distributions is closer to a normal distribution as compared to the non-log ones.

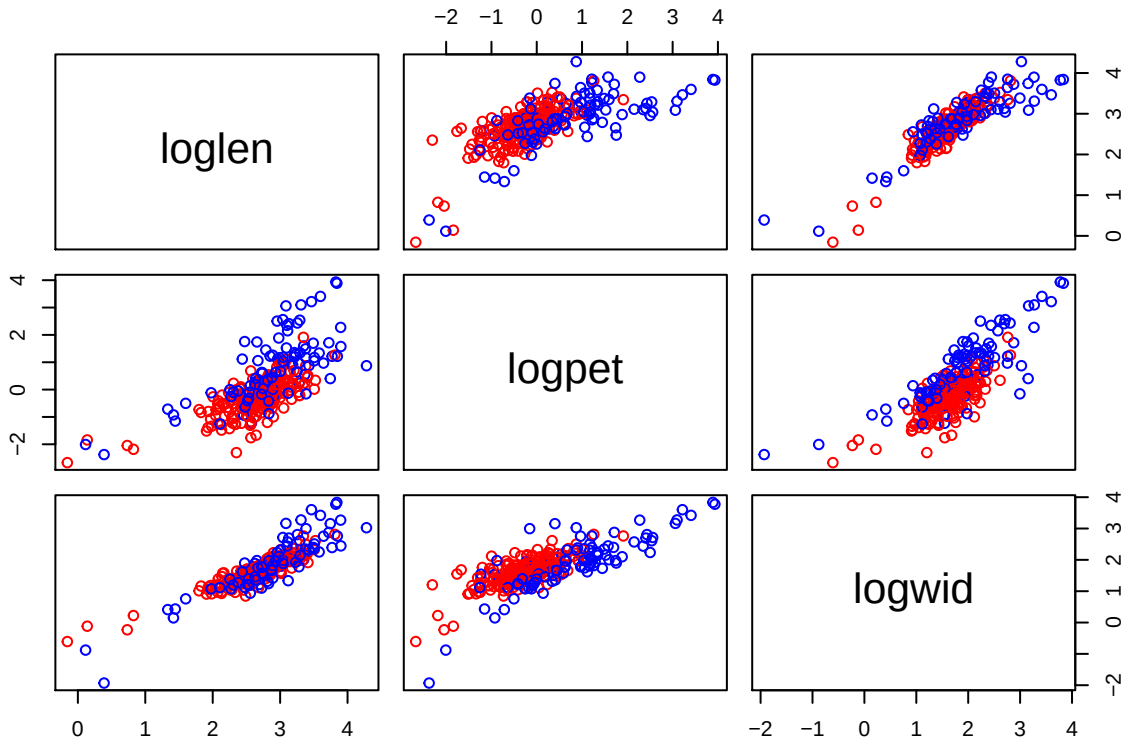
This is similar to the income data where due to the high spread of income, log can work very well.

(b)

```
splomvar <- data %>% dplyr::select(bladelen, petiole, bladewid,arch)
splomlogvar <- data %>% dplyr::select(loglen, logpet, logwid,arch)
cols <- character(nrow(splomvar))
cols[] <- "black"
cols[splomvar[,4] ==1] <- "blue"
cols[splomvar$arch ==0] <- "red"
cols[splomlogvar[,4] ==1] <- "blue"
cols[splomlogvar$arch ==0] <- "red"
pairs(splomvar[,1:3], col=cols)
```




```
pairs(splomlogvar[,1:3], col=cols)
```



The additional information we observed is that the more extreme data at the higher right top end mainly comes from the arch=1 type. The arch=0 type have very feeel high value data and has much lower scatter as compared to the arch==1 type.