Data Analysis and Visualization in R - ggplot

Ikya Jupudy

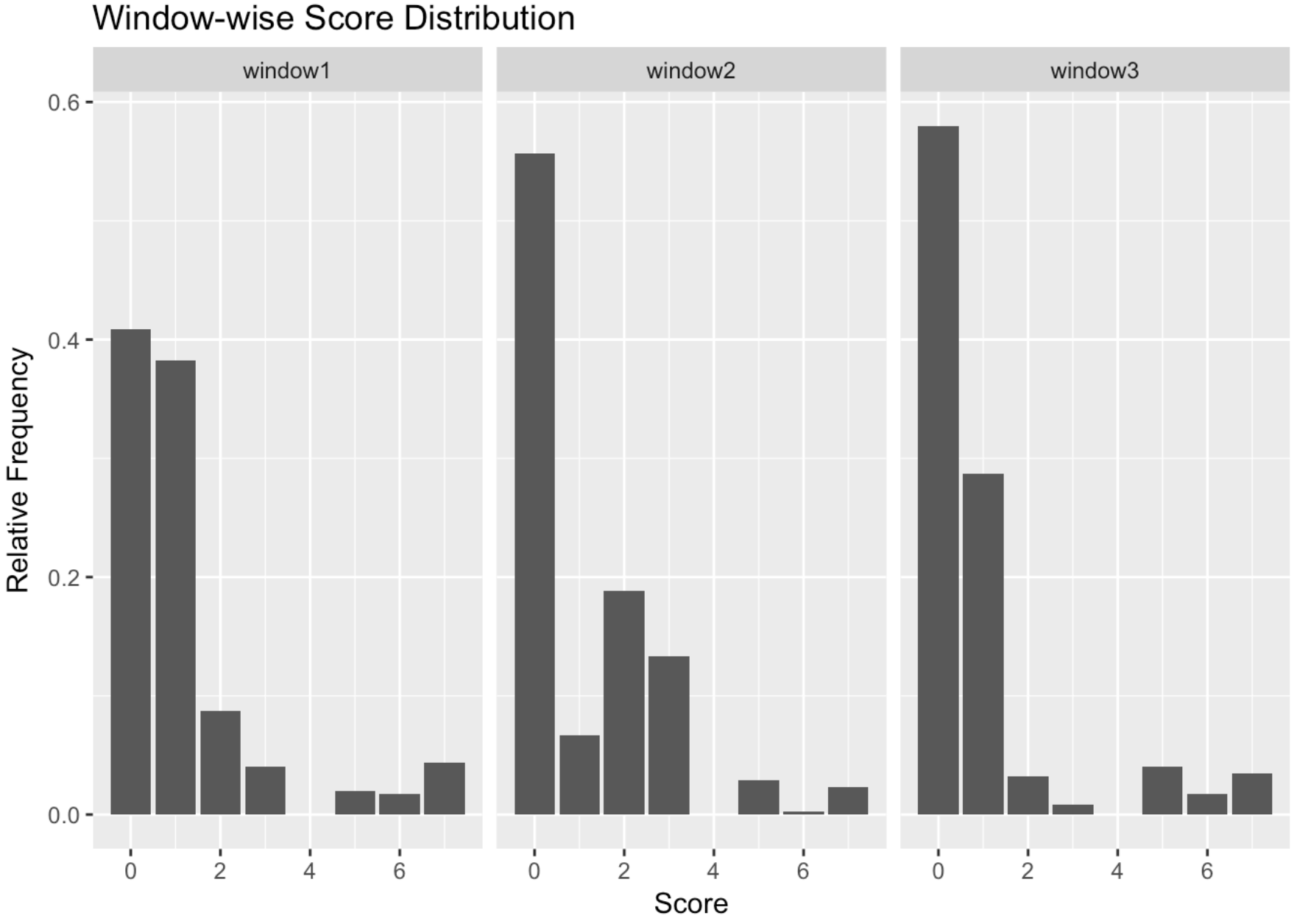
Chapter 4

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('tidyverse')
df<-data.frame(vlt)
dfg<-gather(df,window,Score,-prize,-night)
ggplot(dfg, aes(x=Score)) + geom_bar(aes(y=..prop..))+facet_wrap(~window)+labs(tit
le="Window-wise Score Distribution",y="Relative Frequency")
```



1.The Score for 0 is highest for all the 3 windows.

2.We observe that the first and second symbols occur most frequently accross window 1 and window 3.

3.There is no symbol with representation 4.

4.The 7 symbols are represented by integers 0 to 3 and then 5 to 7.

5.The symbols dont occur across windows with the same probability. For example, the 2nd symbol occurs with much lower probability in the 2nd window as comapred to the first window.

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

[21 points]

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

Preprocessing:- Remove the first column, which is an empty column Then delete the last 54 columns which don't contain any data.

```
df2<-read.delim2('Death2015.txt')
df2 <- df2[ -c(1) ]
df2 <- df2[-c(6897:6940),]
```

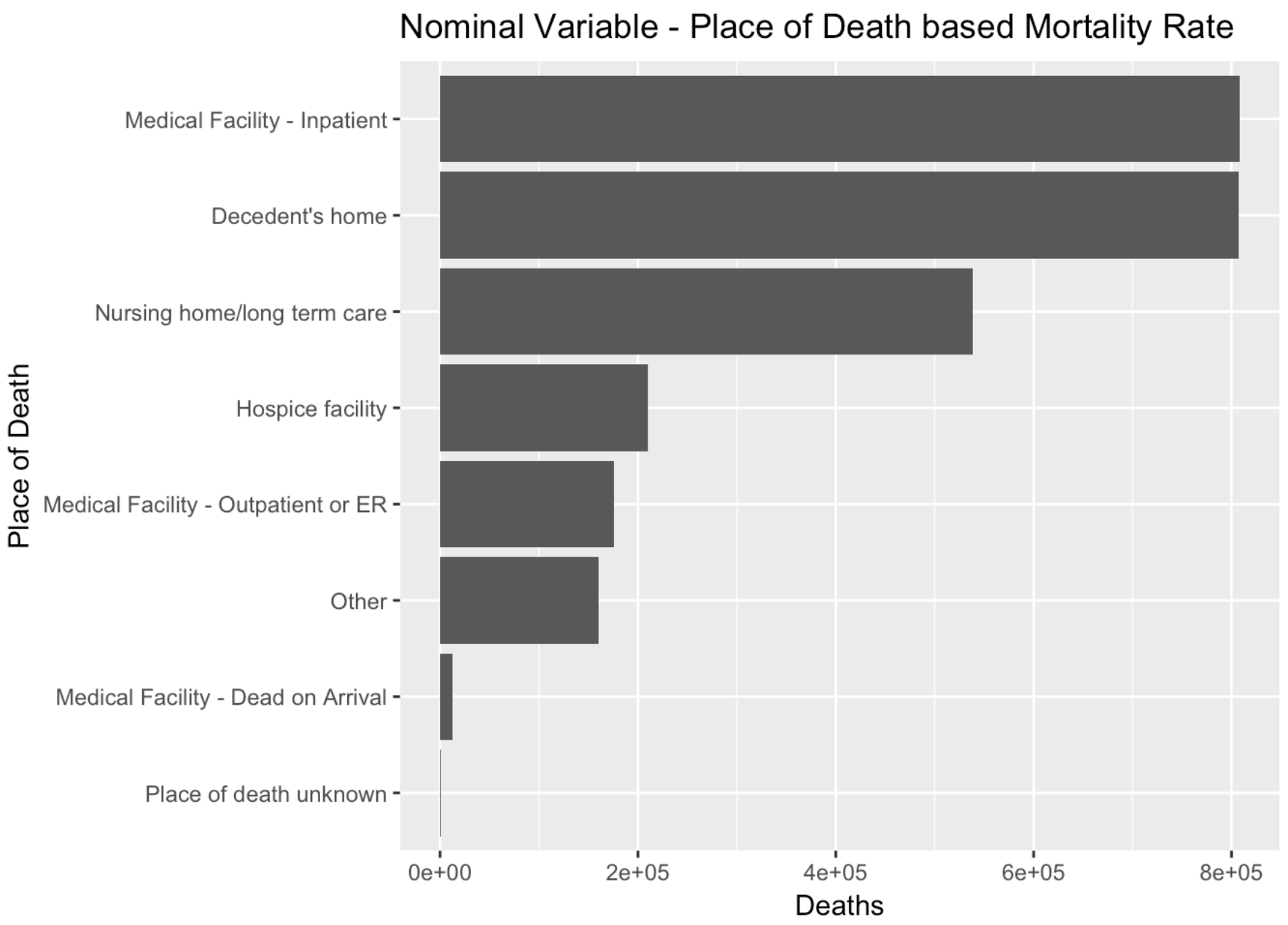
a. For `Place of Death` , `Ten-Year Age Groups` , and `ICD Chapter Code` variables, do the following:

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

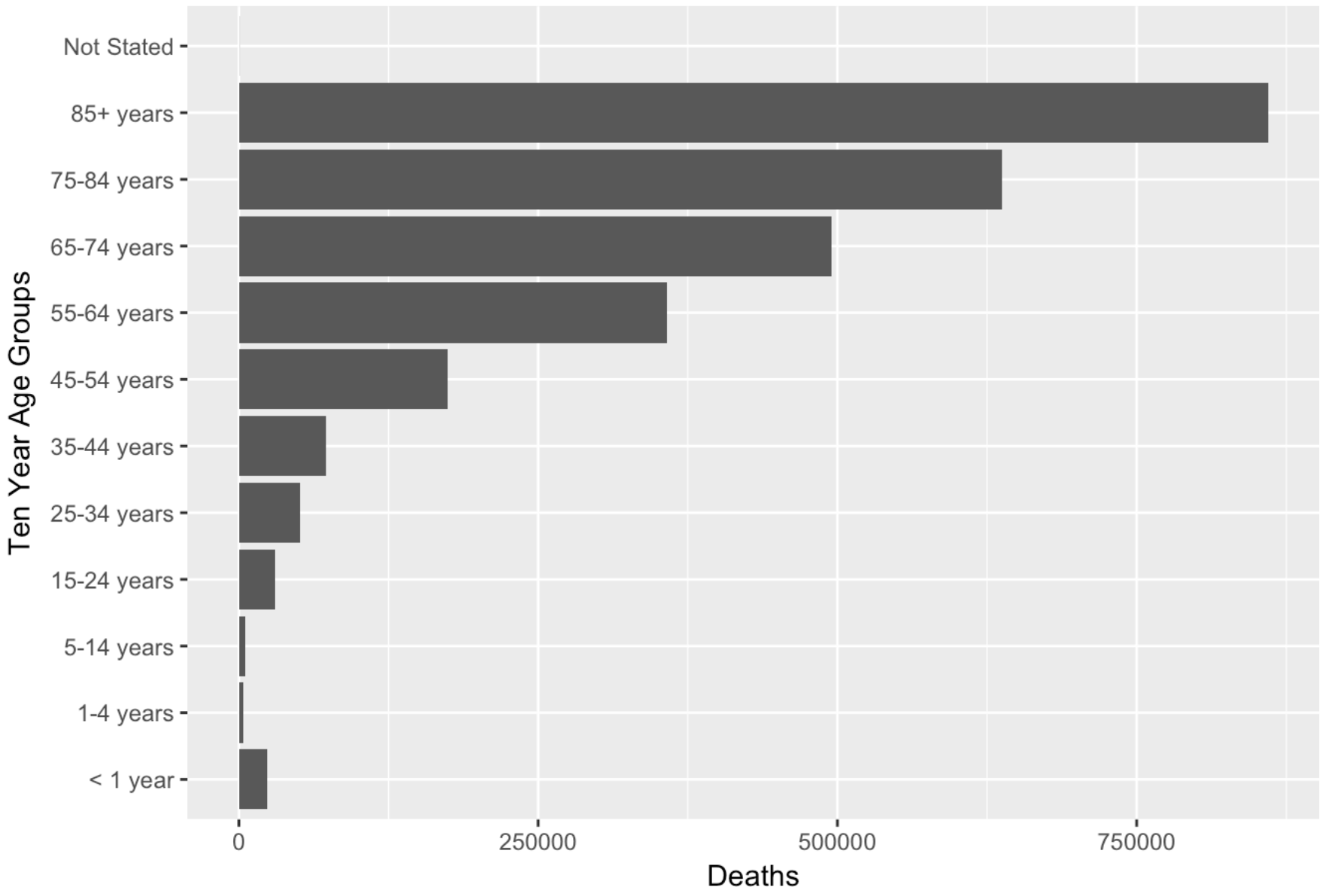
Nominal Variables - Place of Death and ICD Chapter Code

Ordinal Variable - Ten Year Age Groups

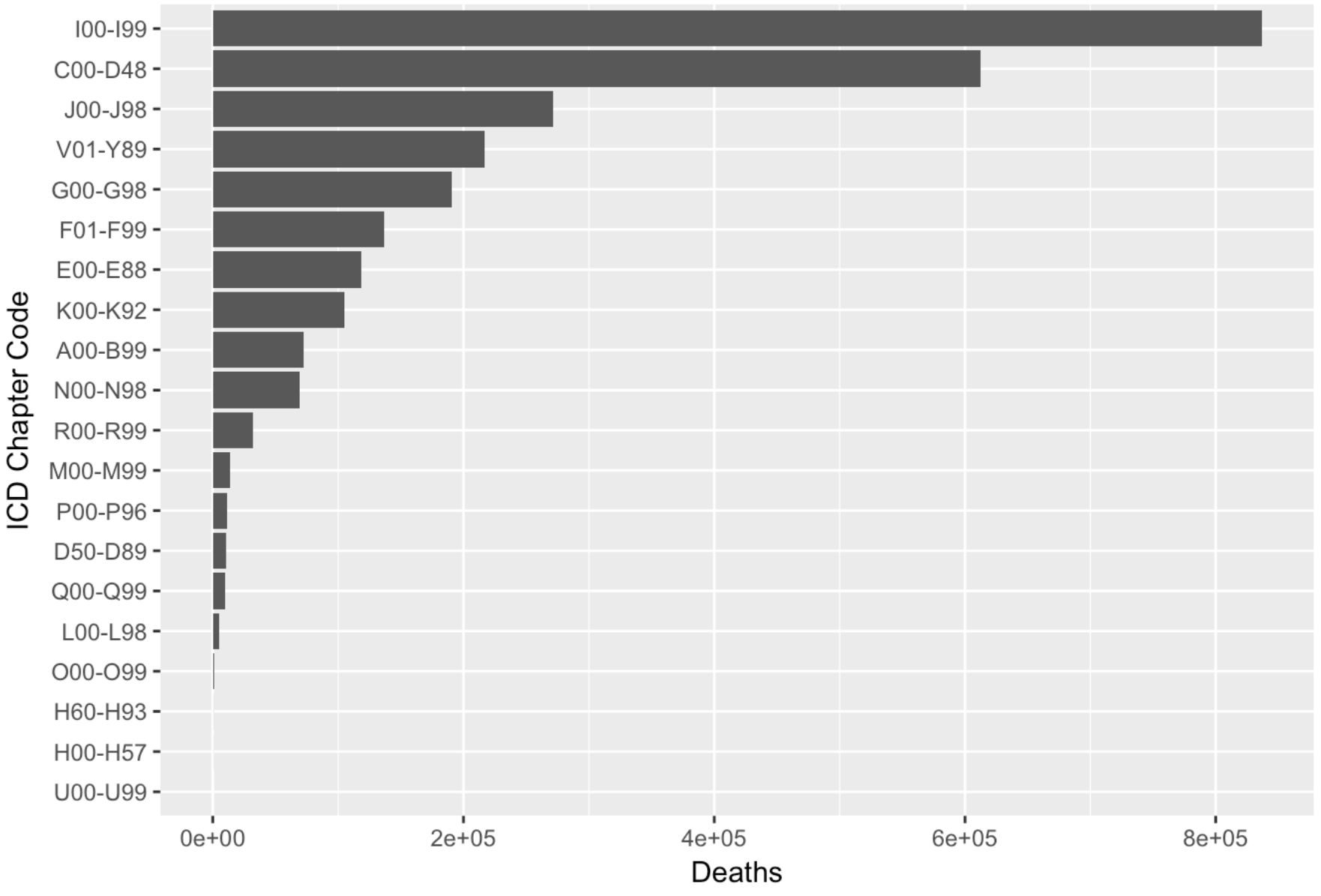
```
ggplot(df2,aes(fct_reorder(Place.of.Death,Deaths,fun=sum)))+
  geom_bar(aes(weight=Deaths))+coord_flip()+labs(title="Nominal Variable - Place o
f Death based Mortality Rate",x="Place of Death",y="Deaths")
```




```
ggplot(df2,aes(fct_relevel(Ten.Year.Age.Groups,"< 1 year","1-4 years","5-14 years"
)))+geom_bar(aes(weight=Deaths))+coord_flip()+labs(title="Ordinal Variable - Ten-Y
ear Age Groups based Mortality Rate",x="Ten Year Age Groups",y="Deaths")
```

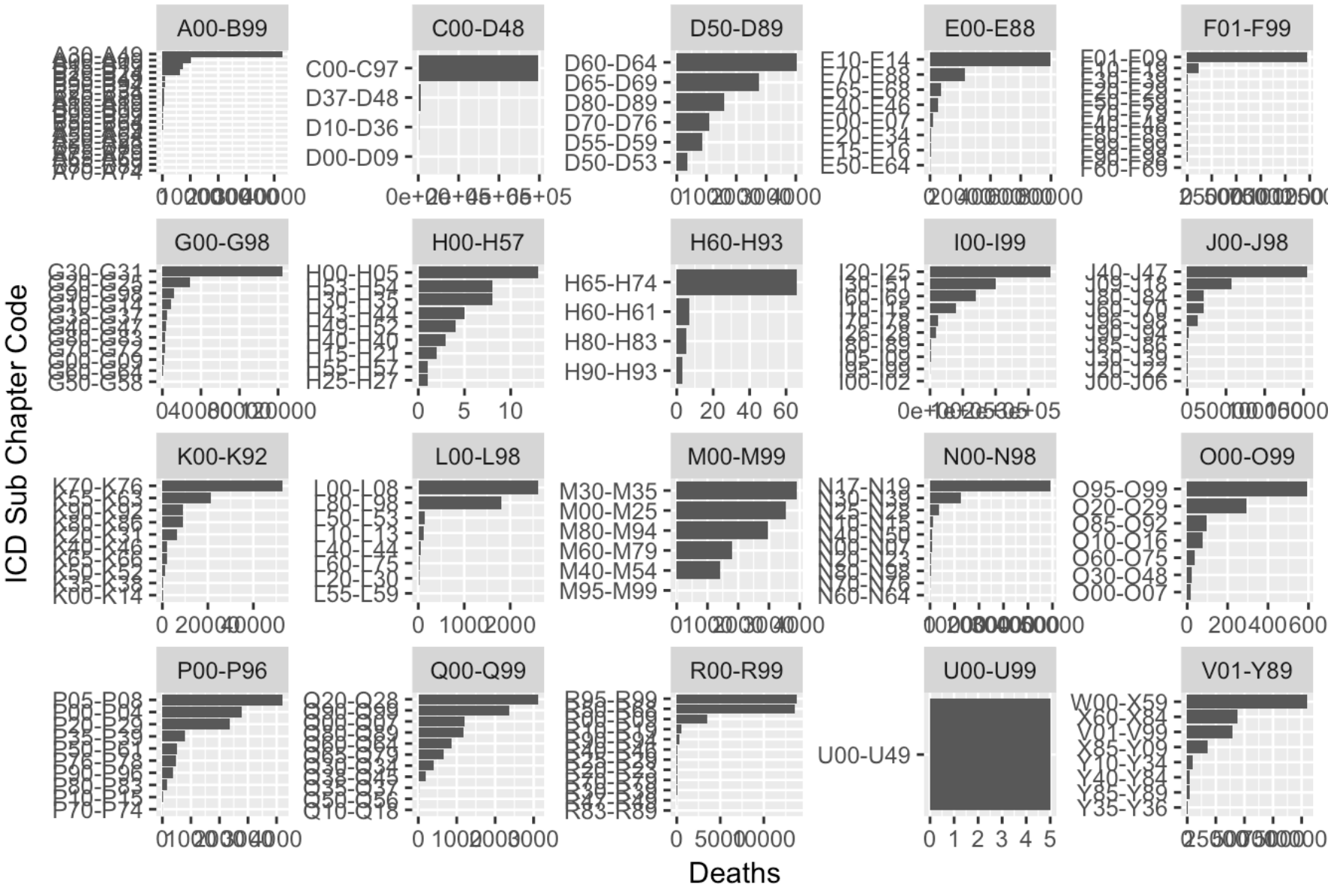



```
ggplot(df2,aes(fct_reorder(ICD.Chapter.Code,Deaths,fun=sum)))+
  geom_bar(aes(weight=Deaths))+coord_flip()+labs(title="Nominal Variable - ICD Cha
pter Code based Mortality Rate",x="ICD Chapter Code",y="Deaths")
```



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.

```
ggplot(df2,aes(fct_reorder(ICD.Sub.Chapter.Code,Deaths,fun=sum)))+
  geom_bar(aes(weight=Deaths))+facet_wrap(~ICD.Chapter.Code,scales="free")+
  coord_flip()+labs(title="Deaths by ICD Sub Chapter for each ICD Chapter",x="ICD
Sub Chapter Code",y="Deaths")
```



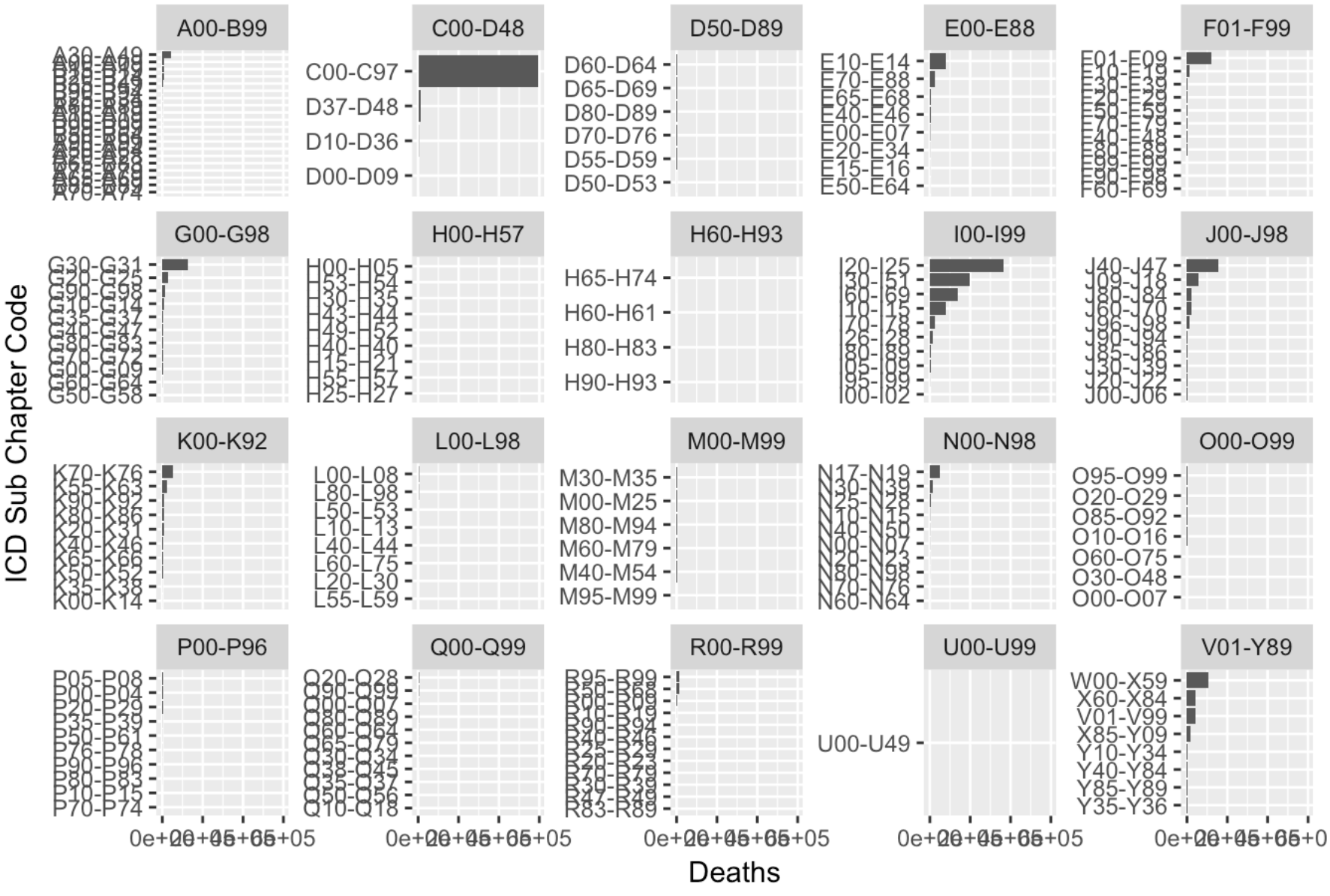
Notable features are:-

1. All the subgraphs have different scales which makes it difficult to compare the subgraphs amongst themselves.

2.Some of the Chapter codes have larger number of Subchapter codes than the others.

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)?

```
ggplot(df2,aes(fct_reorder(ICD.Sub.Chapter.Code,Deaths,fun=sum)))+
  geom_bar(aes(weight=Deaths))+facet_wrap(~ICD.Chapter.Code,scales="free_y")+
  coord_flip()+labs(title="Deaths by ICD Sub Chapter for each ICD Chapter",x="ICD
Sub Chapter Code",y="Deaths")
```

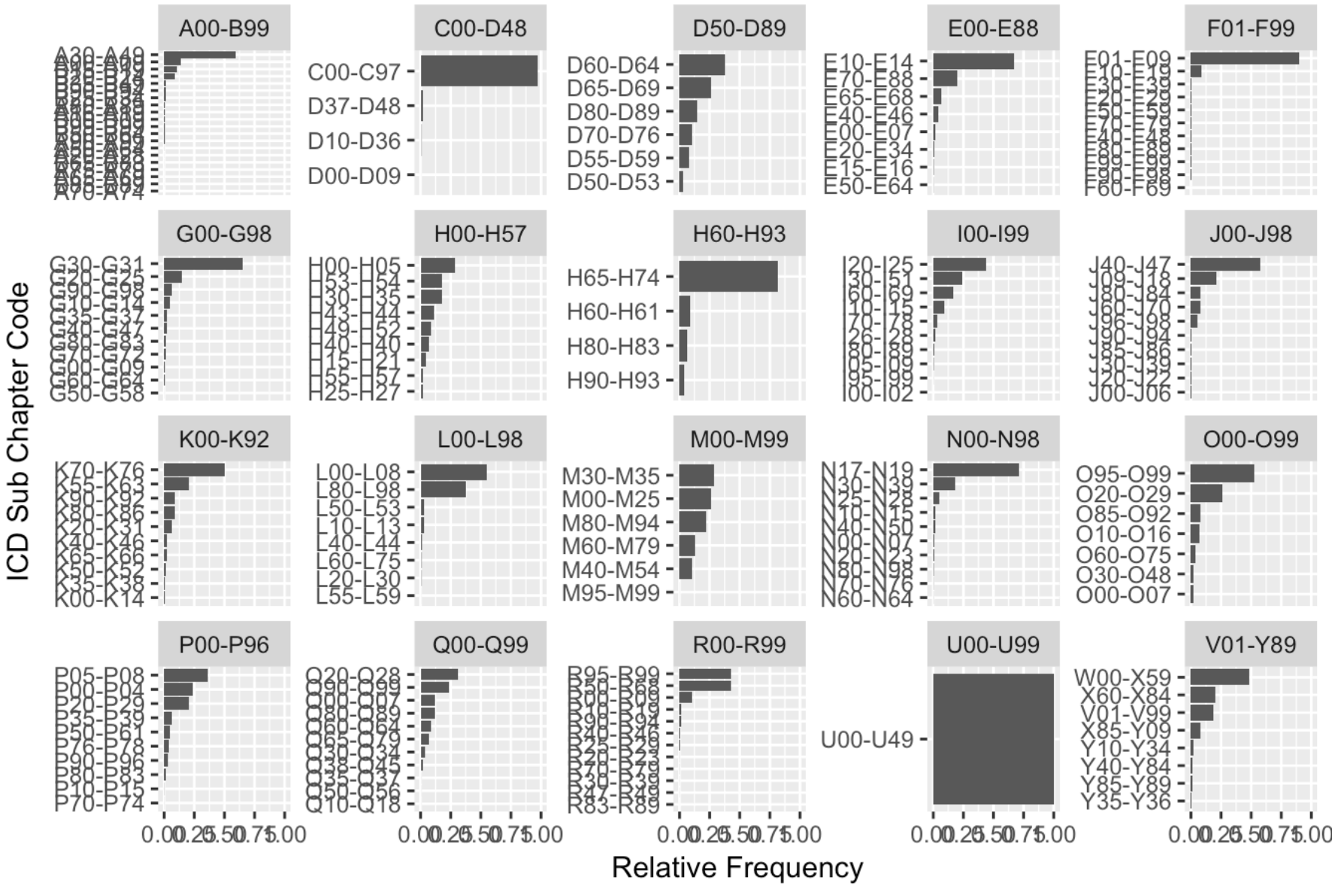


This graph has the same x-axis for all the graphs, which makes it feasible to compare the frequency counts for the different graphs, which was not possible to observe in the previous graph.

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?

```
df22 <- df2 %>%
  group_by(ICD.Chapter.Code,ICD.Sub.Chapter.Code) %>%
  summarise(count= sum(Deaths, na.rm = TRUE)) %>%
  mutate(perc = count/sum(count))

ggplot(df22, aes(fct_reorder(ICD.Sub.Chapter.Code,perc,fun = sum))) +
  geom_bar(aes(weight = perc))+
  facet_wrap(~ICD.Chapter.Code, scales='free_y')+
  xlab("ICD Sub Chapter Code")+
  ylab("Deaths")+
  coord_flip()+labs(title="Deaths by ICD Sub Chapter for each ICD Chapter",x="ICD
Sub Chapter Code",y="Relative Frequency")
```

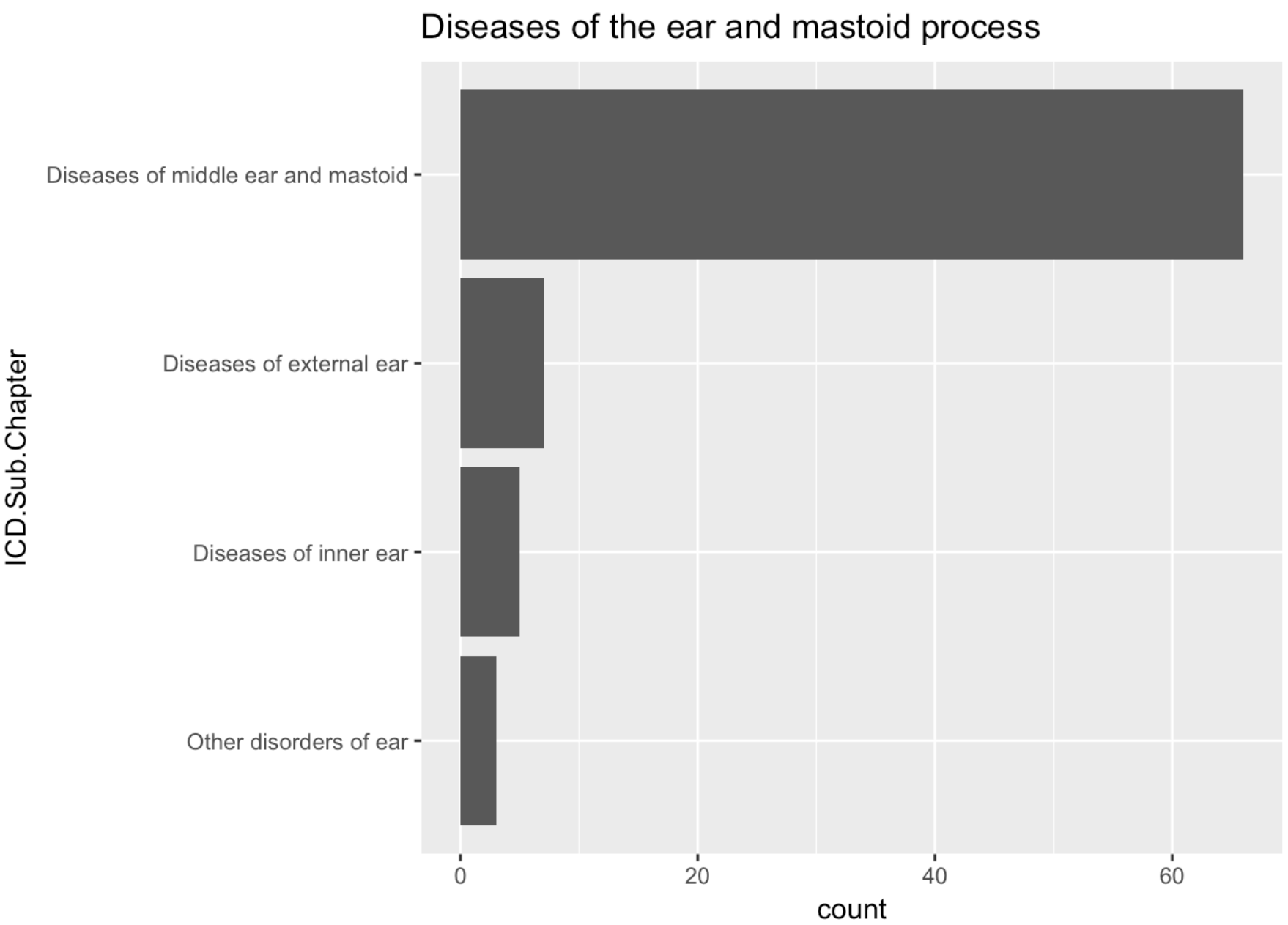


- This graph shows the relative frequencies. i.e. the probability of occurence of a subchapter for a given chapter. Thus for each chapter the probabilities sum to 1. Now its easy to estimate the percentage of data covered by each sub chapter in each respective chapter.

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.

```
df2subset <- subset(df2, ICD.Chapter == ICD.Chapter[32] )

ggplot(df2subset, aes(fct_reorder(ICD.Sub.Chapter,Deaths,fun = sum))) +
  geom_bar(aes(weight = Deaths))+
  coord_flip()+ ggtitle(df2$ICD.Chapter[32])+
  xlab("ICD.Sub.Chapter")
```



This datatype is nominal. In this graph, I plotted the Deaths due to Diseases of the ear and the mastoid process. The different sub categories contributing to this category of Death are Diseases of middle ear and mastoid, Diseases of external ear, Diseases of inner ear and Other disorders.

Maximum number of Deaths are due to Diseases of middle ear and mastoid compared to the other sub-categories.

3. Detailed Mortality, questions about the data

[6 points]

Cite your sources with links.

a. Who is included in the death counts?

Source:- https://wonder.cdc.gov/wonder/help/ucd.html (https://wonder.cdc.gov/wonder/help/ucd.html) The death counts in the data represent deaths that occurred in the 50 United States and the district of Columbia. The data 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.

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

The data query was accessed at http://wonder.cdc.gov/ucd-icd10.html (http://wonder.cdc.gov/ucd-icd10.html) on 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.

Source:-https://en.wikipedia.org/wiki/ICD-10 (https://en.wikipedia.org/wiki/ICD-10) ICD - International Classification of Diseases. The deaths in 2015 are classified using ICD-10 version. Other countries that use the ICD are Russia, South Africa, China, Canada, Germany and many more

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

Source:- https://en.wikipedia.org/wiki/National_Center_for_Health_Statistics (https://en.wikipedia.org/wiki/National_Center_for_Health_Statistics)

Mortality data is collected by the National center for Health Statistics which is a part of CDC (center for disease control and prevention). Headquarters at University Town Center in Hyattsville, Maryland.

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

Source:- https://wonder.cdc.gov/wonder/help/ucd.html (https://wonder.cdc.gov/wonder/help/ucd.html)

Mortality data from the death certificates are coded by the states and provided to NCHS through the Vital Statistics Cooperative Program or coded by NCHS from copies of the original death certificates provided to NCHS by the State registration offices. The documentation mentions that the data with deaths < 20 are unreliable. Therefore, the unreliable death ratio is 0.776% and the reliable death is 99.224%

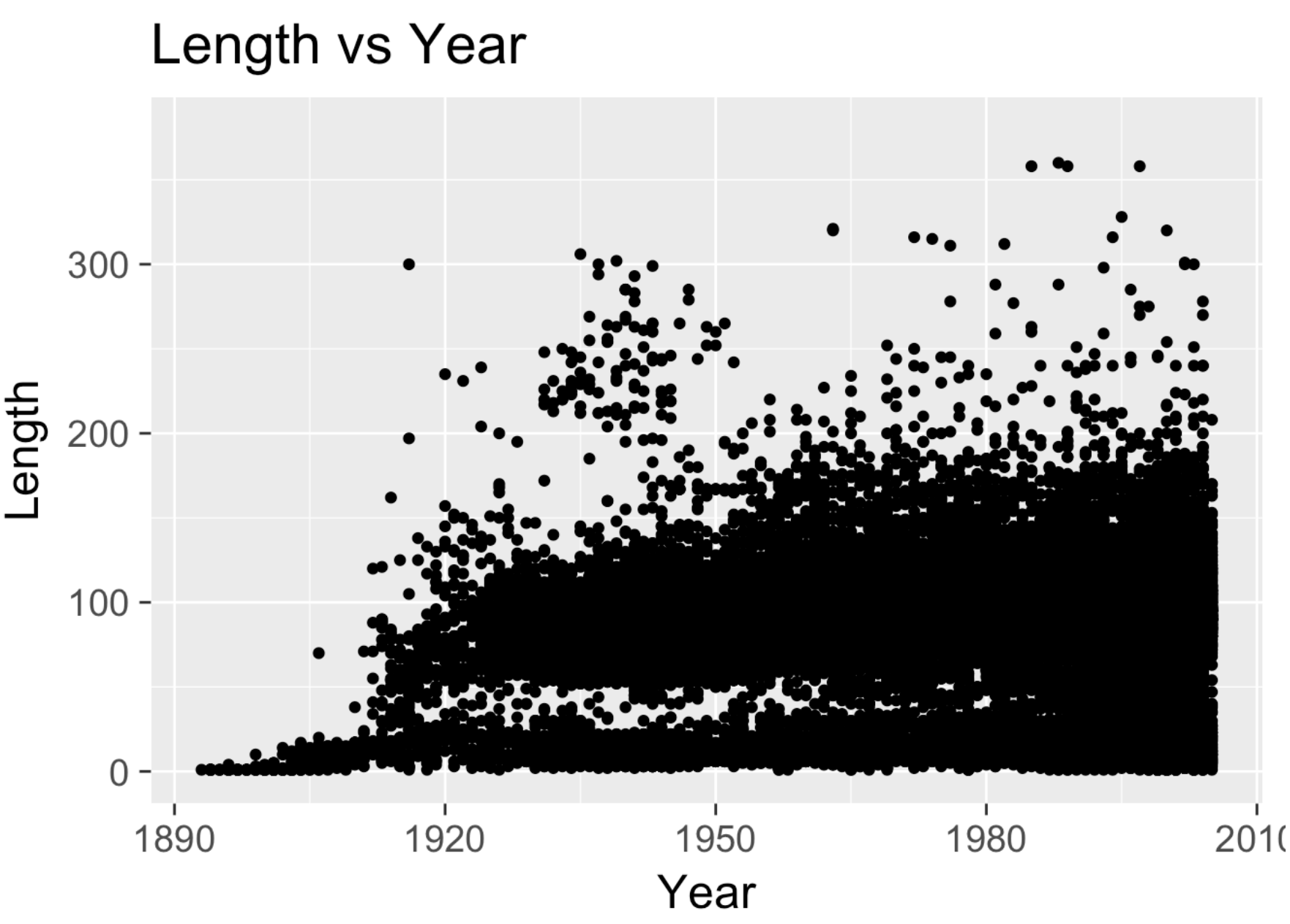
Chapter 5

1. Movie ratings

[12 points]

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

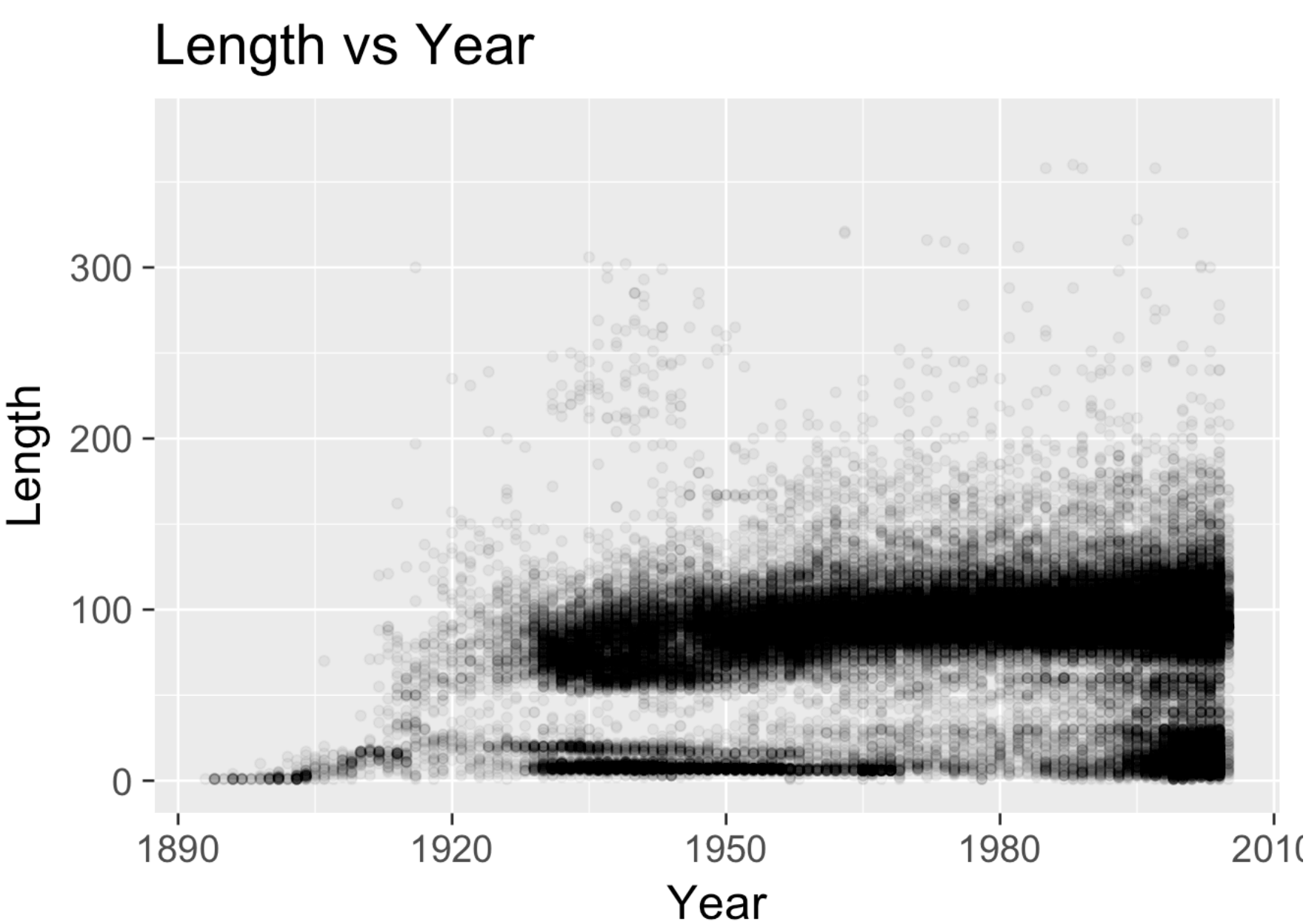
```
library(dplyr)
library(ggplot2movies)
ggplot(movies,aes(year,length))+geom_point()+theme_grey(18)+ylim(0,380)+
  labs(title="Length vs Year",x="Year",y="Length")
```



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

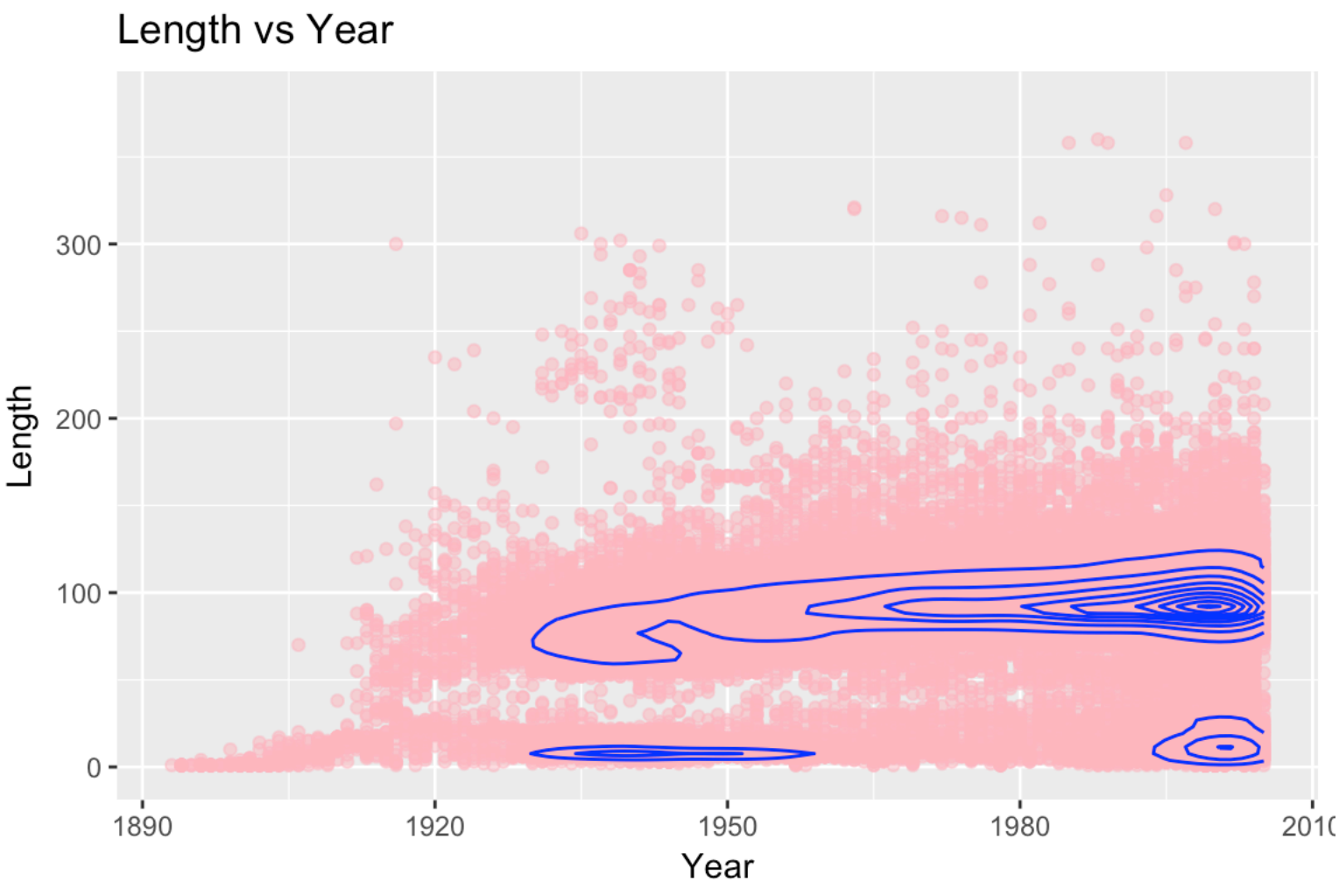
a. Points with alpha blending

```
library(dplyr)
library(ggplot2movies)
ggplot(movies,aes(year,length))+geom_point(alpha=0.05)+theme_grey(18)+ylim(0,380)+
  labs(title="Length vs Year",x="Year",y="Length")
```



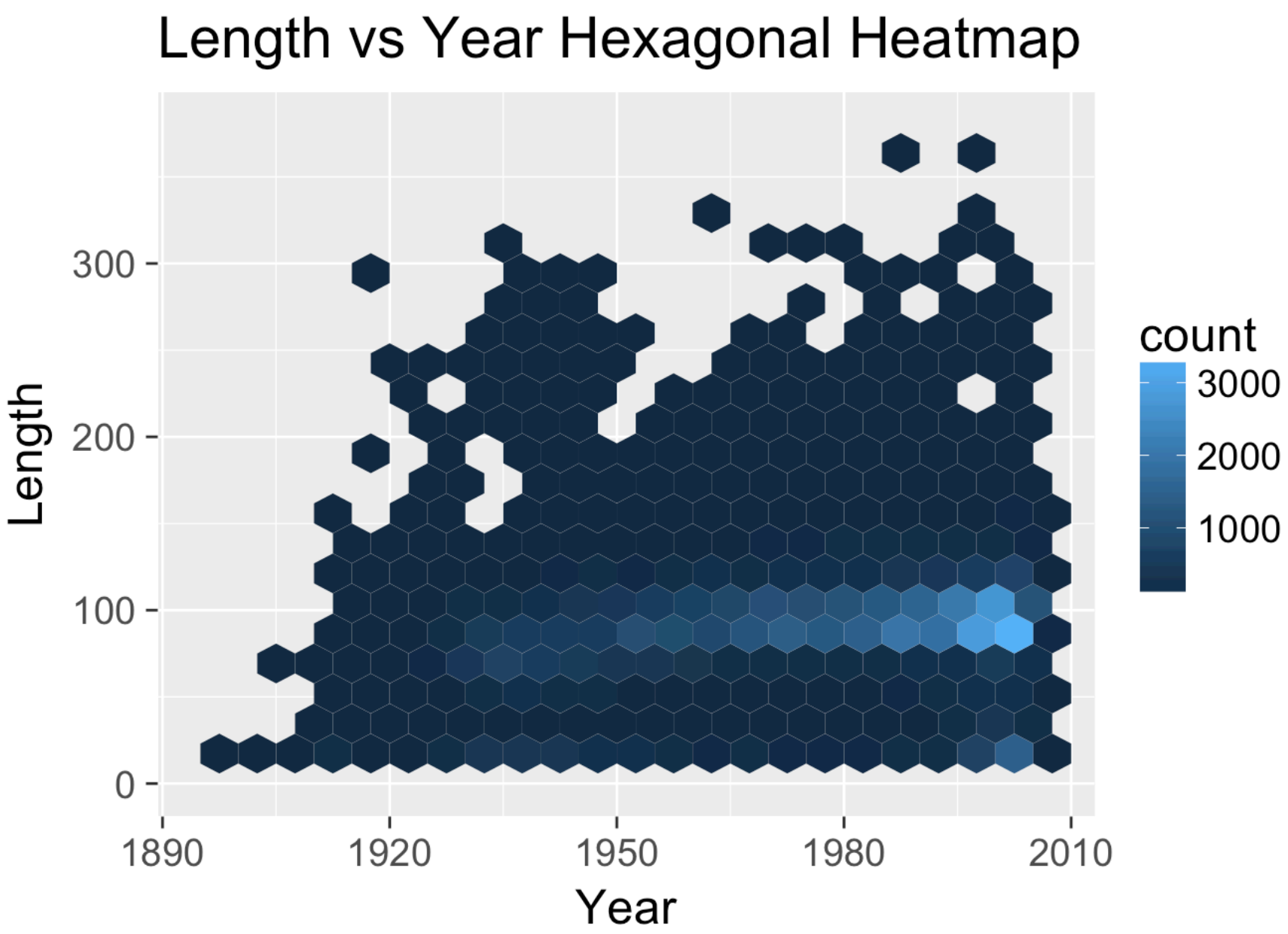
(b) Points with alpha blending + density estimate contour lines

```
ggplot(movies,aes(year,length))+geom_point(color='lightpink',alpha=0.5)+
  geom_density_2d(color='blue')+ylim(0,380)+
  labs(title="Length vs Year",x="Year",y="Length")
```



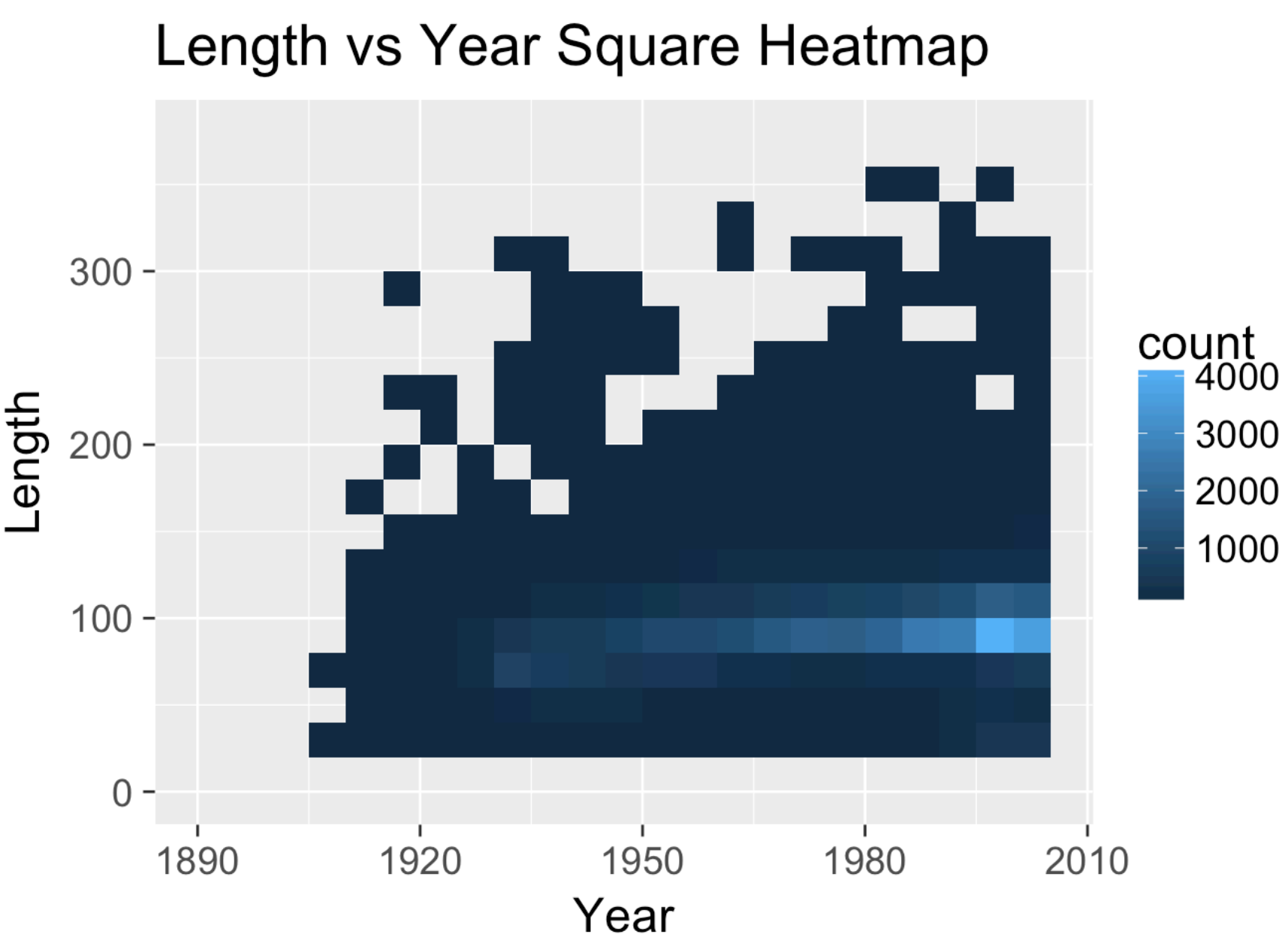
c. Hexagonal heatmap of bin counts

```
g<-ggplot(movies,aes(year,length))+theme_grey(18)+ylim(0,380)
g+geom_hex(binwidth = c(5, 20))+labs(title="Length vs Year Hexagonal Heatmap",x="Year",y="Length")
```



d. Square heatmap of bin counts

```
g<-ggplot(movies,aes(year,length))+theme_grey(18)+ylim(0,380)
g+geom_bin2d(binwidth=c(5,20))+
  labs(title="Length vs Year Square Heatmap",x="Year",y="Length")
```



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.

There are many insights that can be obtained from this plot:-

1.From the hex and square plots we see a distinct line at the 90 minute mark which supports our assumption about average feature length film???s length being around 90 mins

2.There are 2 distinct concentrations of film lengths corresponding to short films and regular feature length films in the contour plots and the seperate blobs for each category.

3.Short films seem to exist more during the period before 1920.Feature length films start to appear more frequently after 1920.

4.The number of films increases over time.And there are more films with intermediate lengths, as is evident by the gap between feature length and short films.

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

1. The contour plot tells us that most of the feature length films from the 70s to the 2000s are between 60 - 120 mins, while short films are between ~1-25 mins.The contour plots also point to an isolate concentration of short films in the 1930-1950s

2.From the hex and square heatmaps we notice that there is a greater concentration of 90 minute films in recent years i.e 90s and 2000s as compared to the 70s and 80s. This is not evident from the alpha blending points

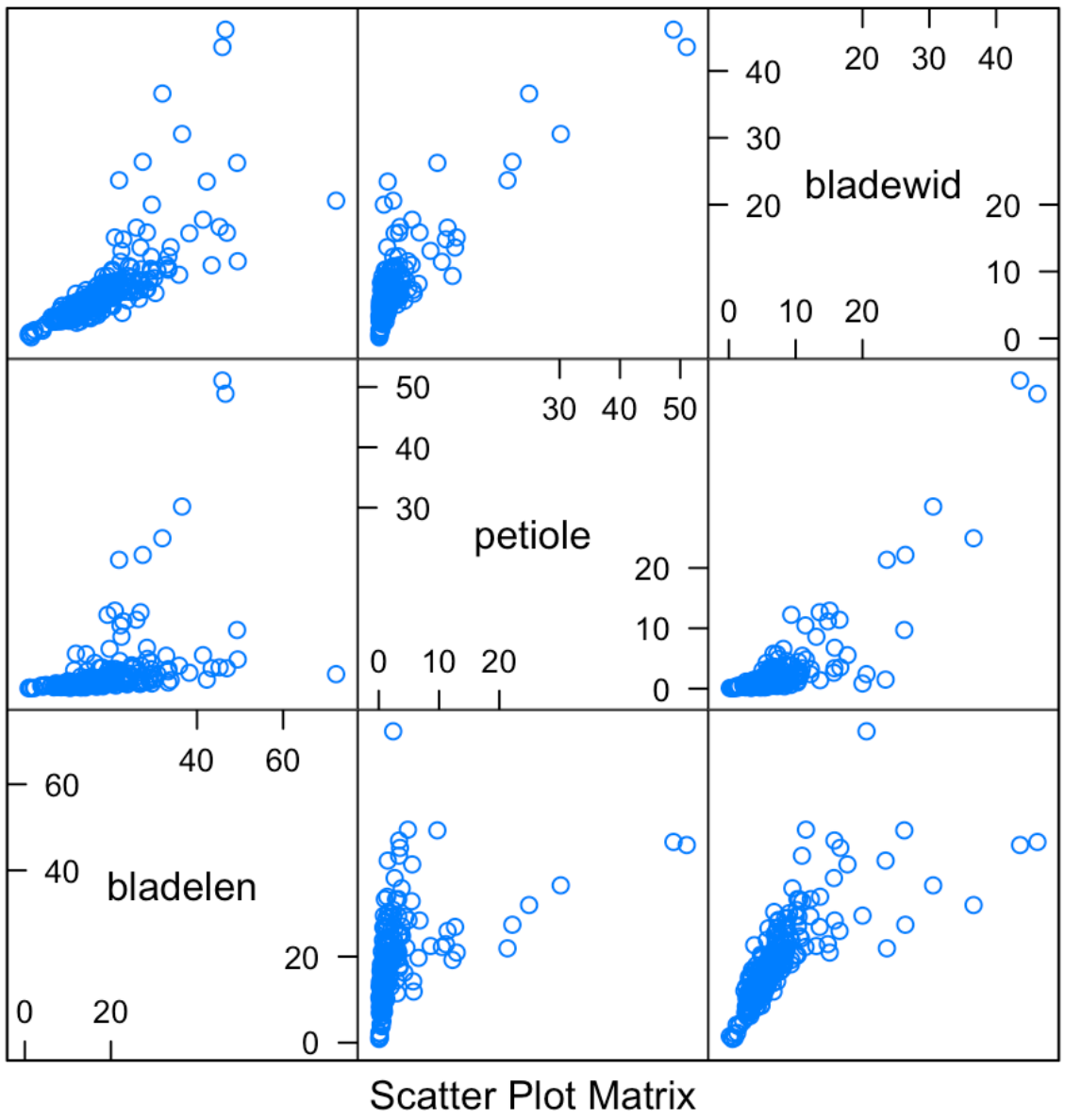
3. Heatmaps indicate that from around 2000s,

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

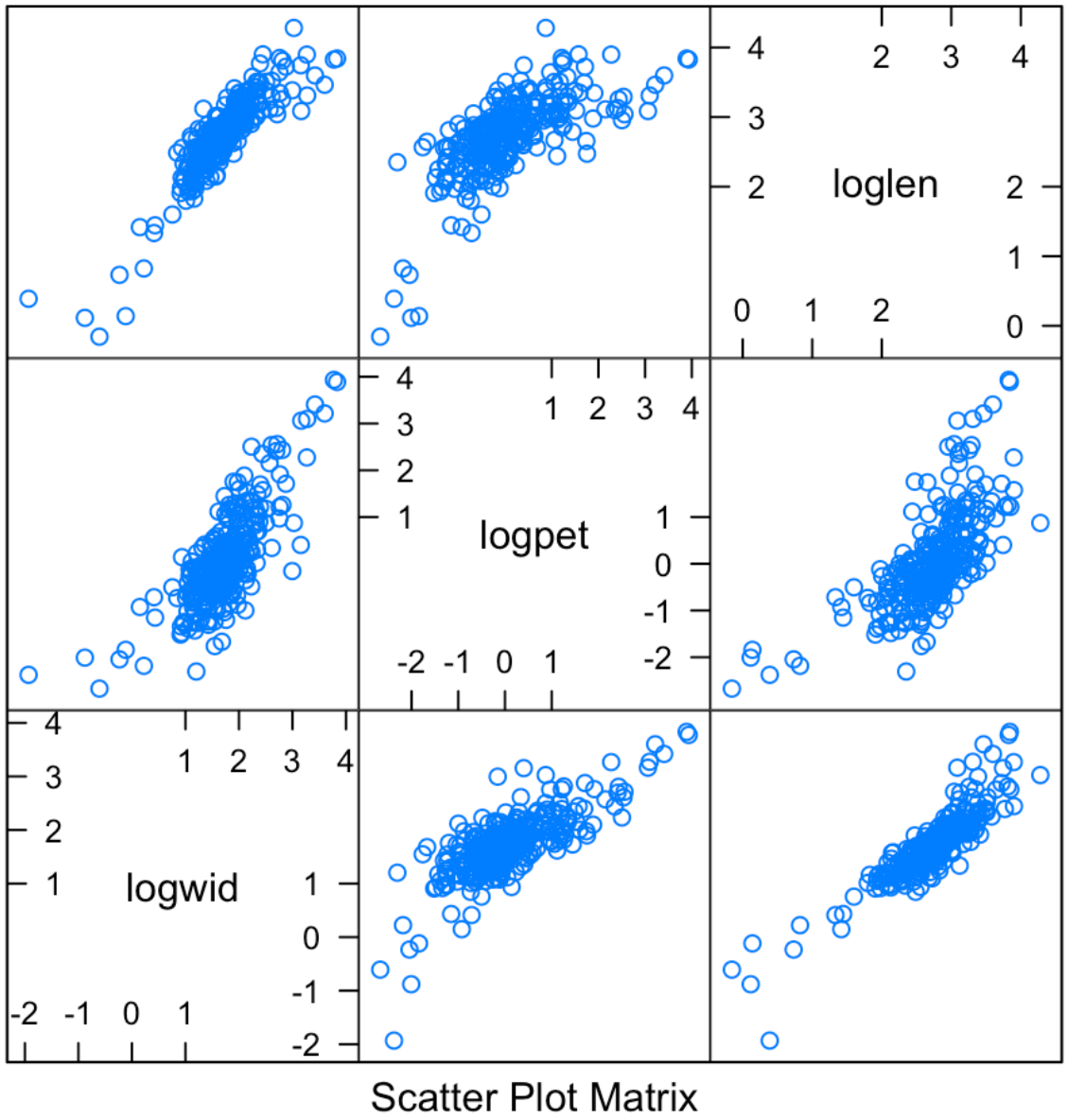
[6 points]

Scatterplot matrices of Plot 1:-Length, Petiole and Width
Plot 2:-Length, Petiole and Width Log Variables

```
ldata<-leafshape
splom(~ldata[1:3])
```



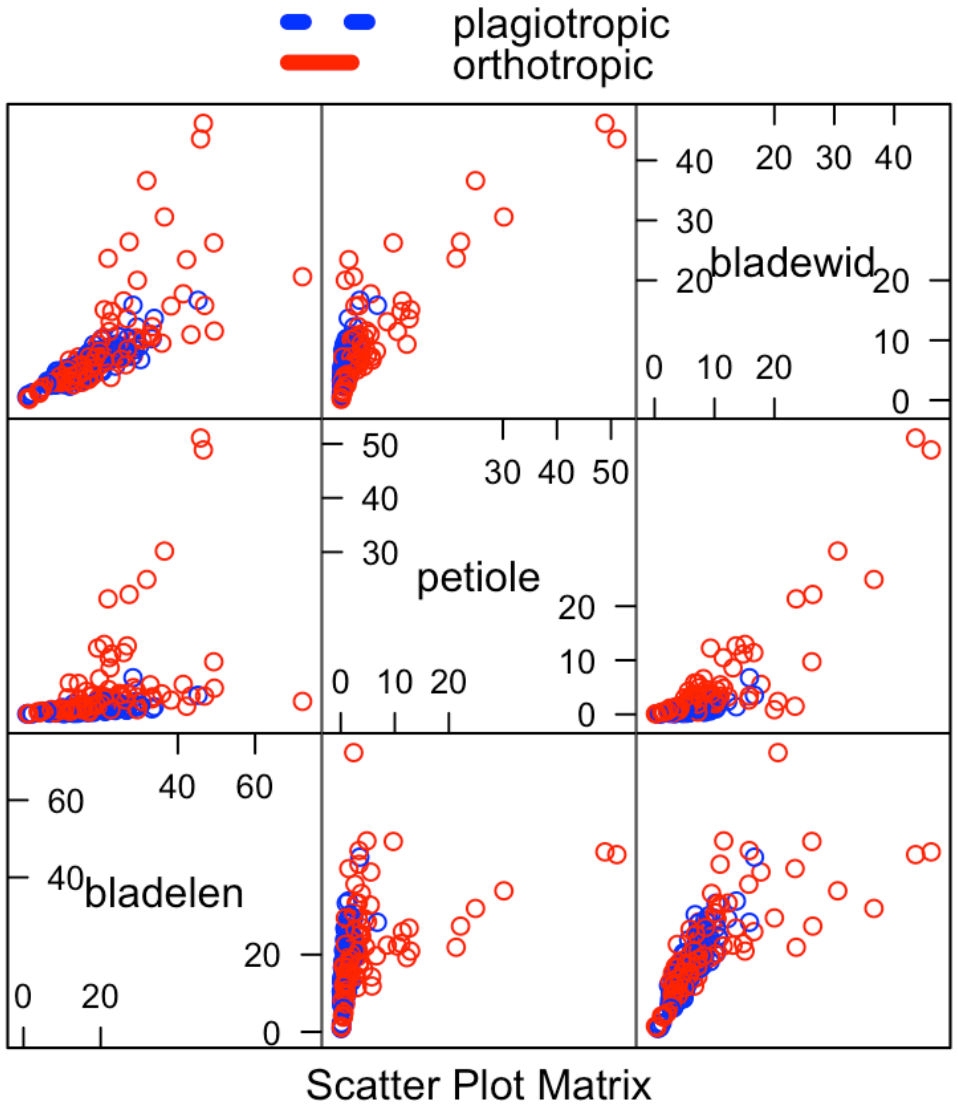

```
splom(~ldata[5:7])
```



Scatter Plot Matrix

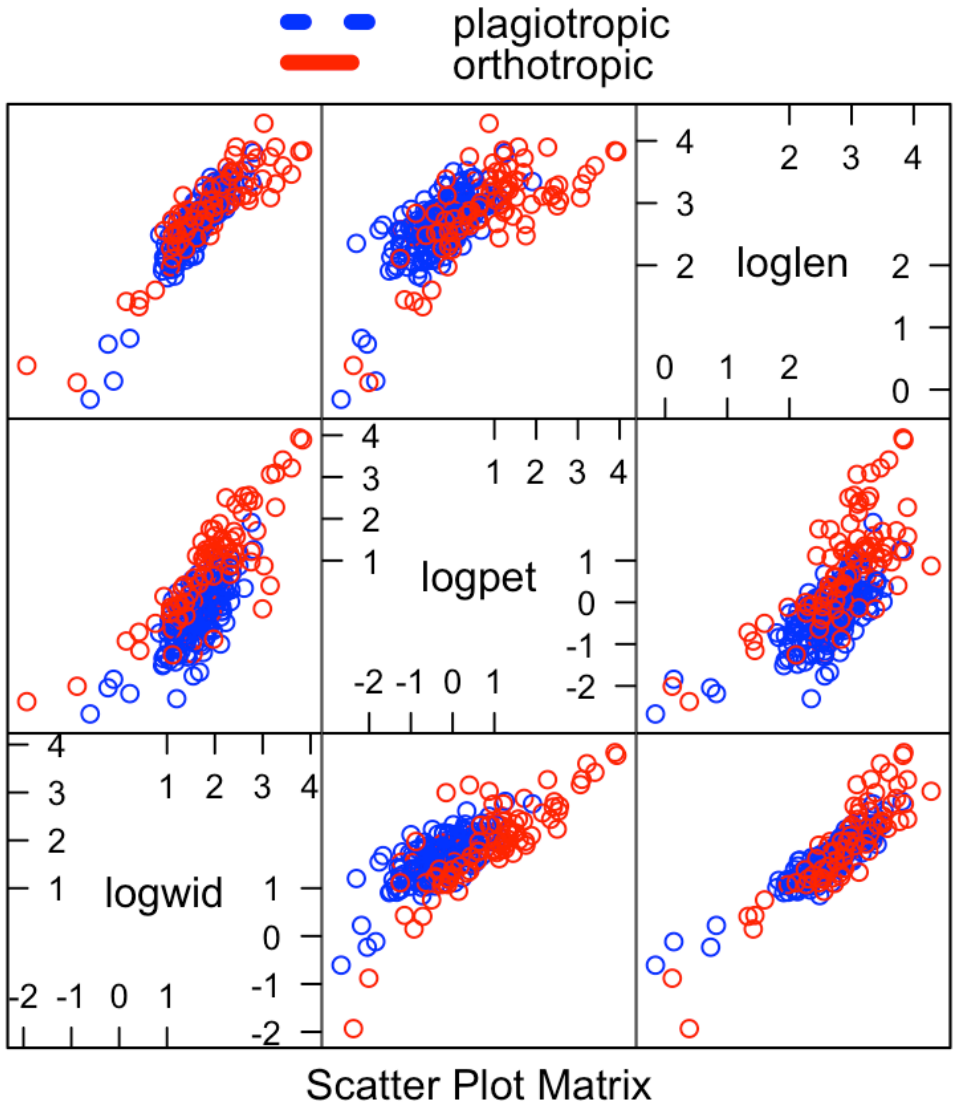
```
splom(~ldata[1:3],col = c("blue","red")[ldata[,8]+1],
      key=list(space="top",title="SPLOM matrix of Length, Petiole and Width by var
iable Arch",
               lines=list(col=c("blue","red"),lty=c(3,2),lwd=6),
         text=list(c("plagiotropic", "orthotropic"))))
```

SPLOM matrix of Length, Petiole and Width by variable Arch




```
splom(~ldata[5:7],col = c("blue","red")[ldata[,8]+1],
      key=list(space="top",title="SPLOM of Length, Petiole and Width Log Variables
by Arch",
               lines=list(col=c("blue","red"),lty=c(3,2),lwd=6),
               text=list(c("plagiotropic", "orthotropic"))))
```

SPLOM of Length, Petiole and Width Log Variables by Arch



From the first set of scatter plot matrix between Length, Width and Petiole variables, most of the data is saturated towards one corner of the plot. The second set of scatter plot matrix between Length, Width and Petiole log variables, the data is more spread out and easy to analyse.

From the next 2 graphs, the scatterplot matrix based on with the log variables look Under the log transformation, the two colors are seperated out better. We can see the distinction between the two different leaf architectures better in this graph The log linearity of blade length and blade width is maintained The Petiole is observed to be smaller in Plagiotrpic