Problem Set 1 Spring 2022

Note: Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class, including choosing appropriate parameters for all graphs. *Do not expect the assignment questions to spell out precisely how the graphs should be drawn. Sometimes guidance will be provided, but the absense of guidance does not mean that all choices are ok.*

Read *Graphical Data Analysis with R*, Ch. 3

1. SATs

[7 points]

Data: *StudentSurvey* in **Lock5withR** package (Remember to add a proper title and labels to every plot.)

#a) Draw multiple horizontal boxplots of `SAT`, by `Year`. What do you observe? (Hint:You can remove all blank and NAs)

```
#install.packages("Lock5withR")
library(Lock5withR)
library(tibble)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
df1<- tibble(StudentSurvey)
```

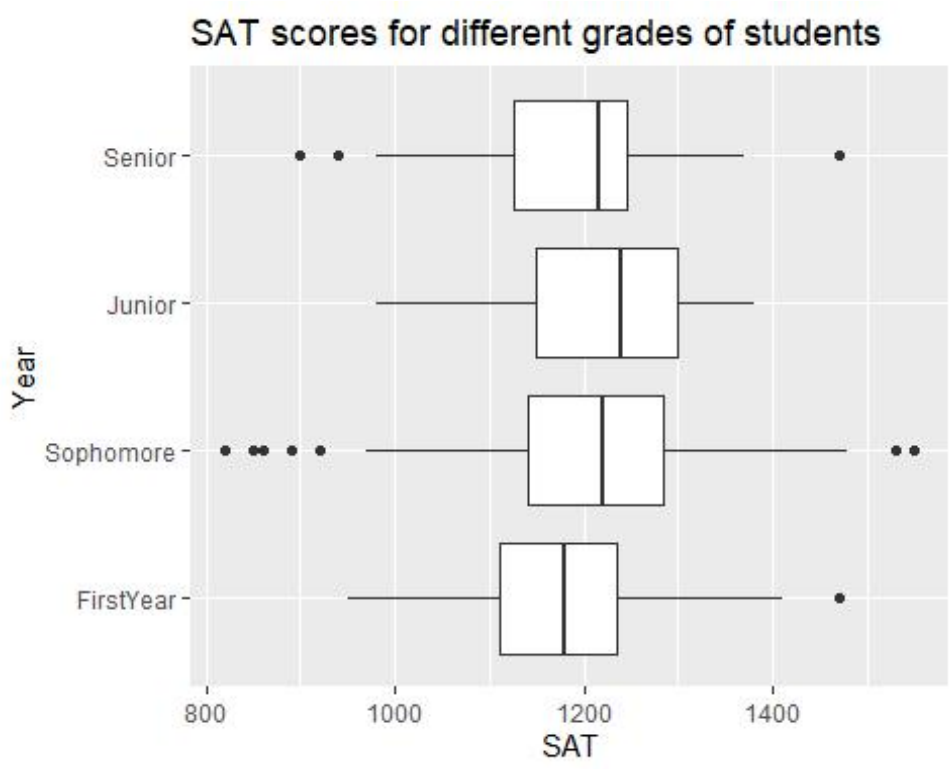
```
typeof(df1$Year)
```

```
## [1] "integer"
```

```
unique(df1$Year)
```

```
## [1] Senior    Sophomore FirstYear Junior
## Levels:  FirstYear Junior Senior Sophomore
```

```
df1 %>%
  mutate( Year = fct_relevel( df1$Year, "FirstYear", "Sophomore", "Junior")) %>%
    na.omit() %>%
  ggplot(aes( x = SAT, y = Year)) +
  geom_boxplot() +
  labs( title = "SAT scores for different grades of students" )
```



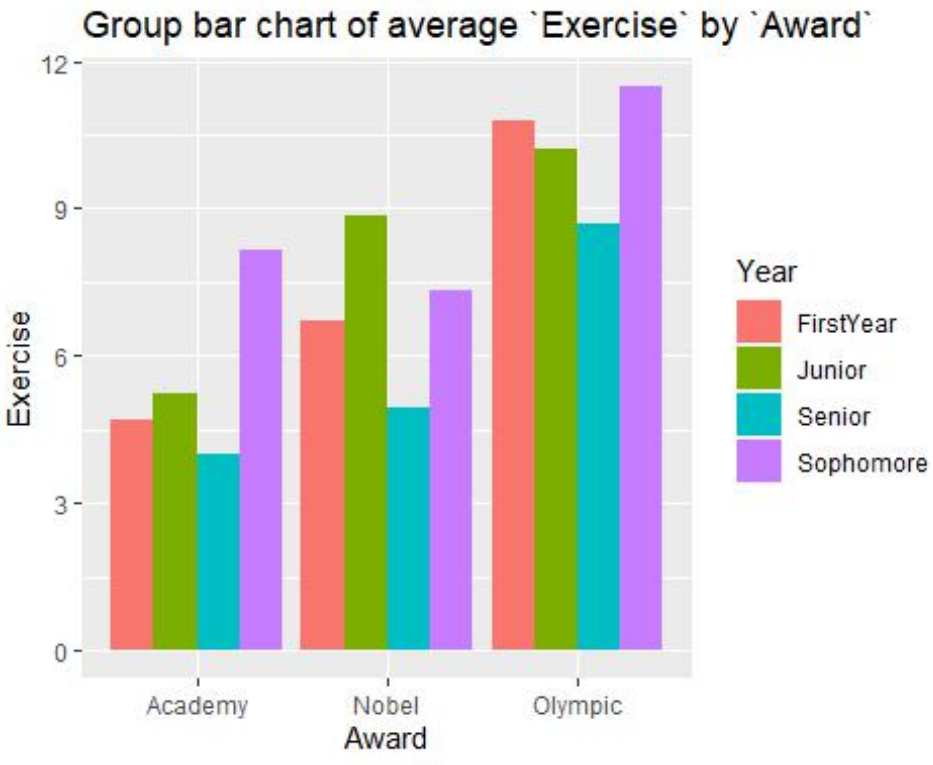
It can be observed that in terms of median grade, junior students are doing better than all other grades, and first year students are doing the worst. Also, sophomore students SAT scores have outliers in two sides of the box, which means that there are sophomore students doing very good or very bad in the SAT tests. On contrary, there is no outlier for Junior students, meaning that junior students are doing well overall without anyone doing super well or getting left behind. For senior students there are 3 outliers, and for first year students there is 1 outlier.

#b) Draw a grouped bar chart of average `Exercise` by `Award` filled with `Year`. (You can ignore NAs.)

```
df1 %>%
  na.omit() %>%
  ggplot(aes( x = Award, y = Exercise, fill = Year)) +
  geom_bar(position = "dodge", stat = "summary", fun.y = "mean")+
  labs( title = "Group bar chart of average `Exercise` by `Award`" )
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```

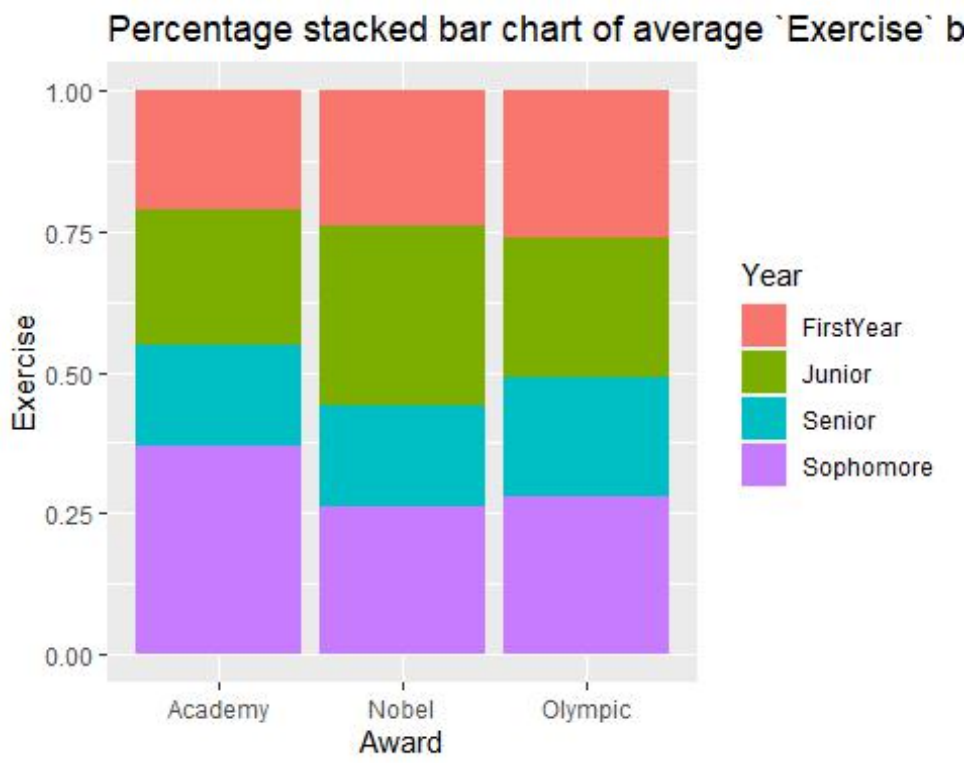


#c) Draw a percentage stacked barchart (each bar = 100%) of average `Exercise` by `Award` filled with `Year`. Compare to the plot in b), which one do you prefer and why?

```
df1 %>%
  na.omit() %>%
  ggplot(aes( x = Award, y = Exercise, fill = Year)) +
 # geom_bar(position="fill", stat="identity") +
  geom_bar(position = "fill", stat = "summary", fun.y = "mean")+
  labs( title = "Percentage stacked bar chart of average `Exercise` by `Award`" )
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```



Which one I would prefer depends on the problem I want to analyse. For the percentage stacked barchart, it could be easier for us to compare for students in the same grade, the difference of average time of exercise they would do when they want to win different awards. And for the boxplot, it is easier to compare the difference of average exercise time for students who want to win different kinds of award.

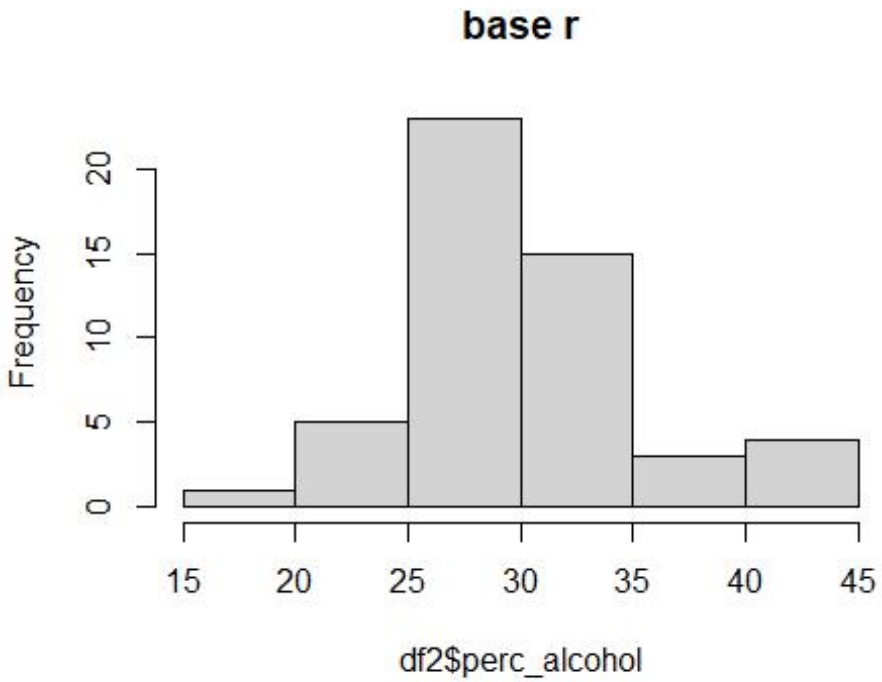
2. Bad Drivers

[7 points]

Data: *bad_drivers* in **fivethirtyeight** package

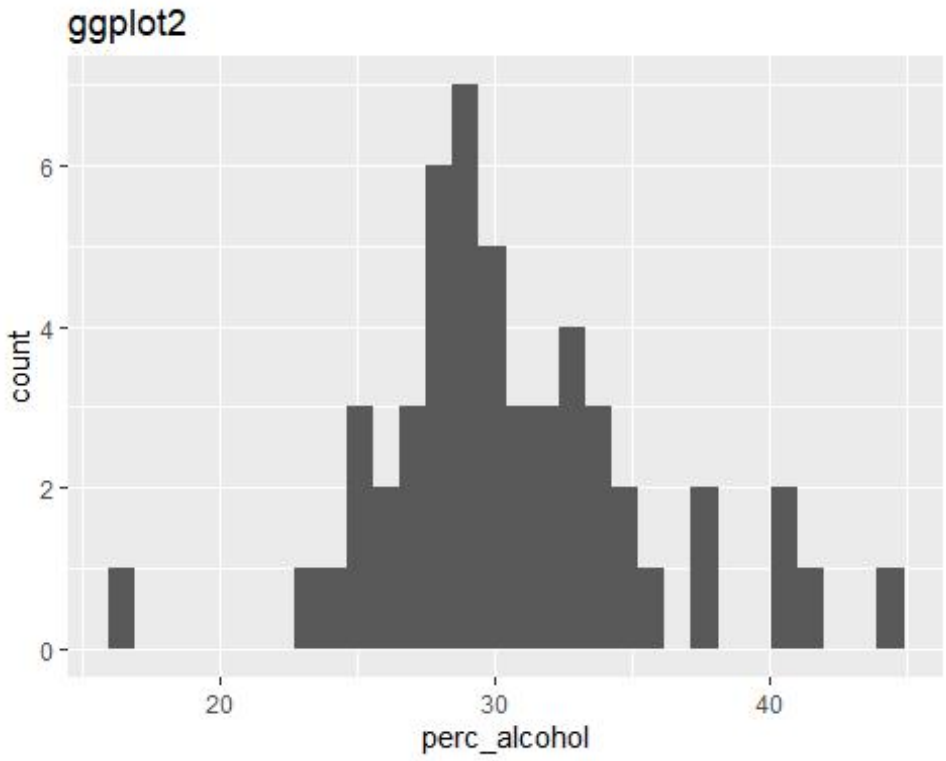
#a) Draw two histograms–one with base R and the other with **ggplot2**–of the variable representing the Percentage of drivers involved in fatal collisions who were alcohol-impaired without setting any parameters. What is the default method each uses to determine the number of bins? (For base R, show the calculation.) Which do you think is a better choice for this dataset and why?

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

## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')

df2<- tibble(bad_drivers)

hist(df2$perc_alcohol, main = "base r")
```




```
df2 %>%
  ggplot( aes(x = perc_alcohol)) +
  geom_histogram() +
  labs( title = "ggplot2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For bins, ggplot2 uses 30 bins by default while base R hist function uses the Sturges method to calculate the number of bins.

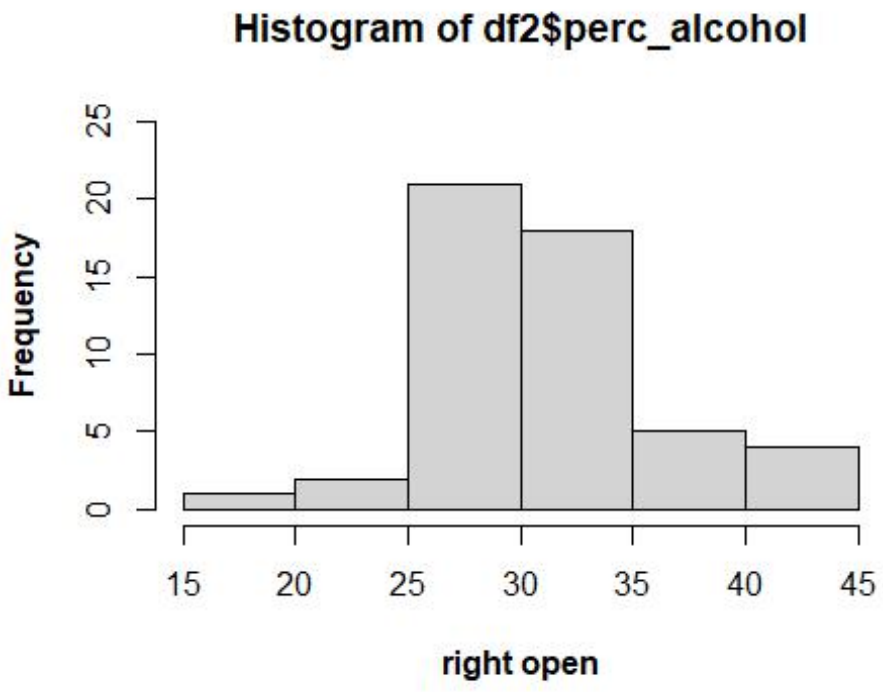
For base r, Sturges' Rule uses a formula to determine the optimal number of bins to use in a histogram: Optimal Bins = $[\log_2 n + 1]$, where n is the total number of observations in the dataset, and the [] means ceiling, which is rounding the number inside up to the nearest integer.

I think the default method of base r is better because it can give a bin that can fit the data better. This can be shown clearly from the two plots above, from the ggplot2 plot we can see there are a lot of gaps between the bars, which means that there is no data in the range. So the default bins for ggplot2 is not very good.

#b) Draw two histograms of the `perc_alcohol` variable with boundaries at multiples of 5, one right closed and one right open. Every boundary should be labeled (15, 20, 25, etc.)

```
hist(df2$perc_alcohol, ylim = c(0, 25), xlab = "right closed", font.lab = 2)
```

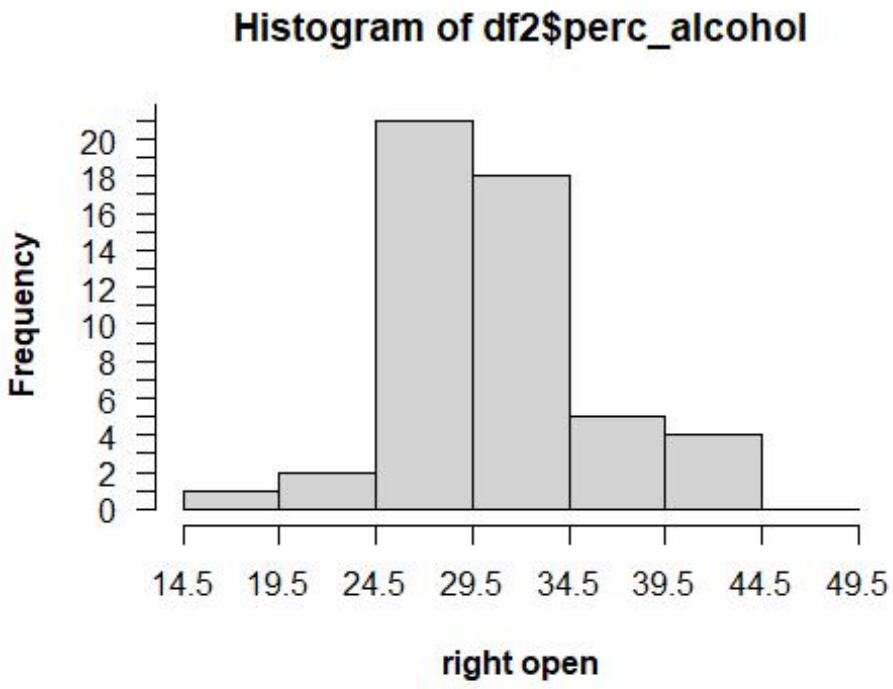

```
hist(df2$perc_alcohol, right = FALSE, ylim = c(0, 25), xlab = "right open", font.lab = 2)
```



#c) Adjust parameters–the same for both–so that the right open and right closed versions become identical. Explain your strategy.

```
hist(df2$perc_alcohol, breaks = seq(14.5, 49.5, 5),
     xlab = "right closed", font.lab = 2, axes = FALSE)
axis(1, at = seq(14.5, 49.5, 5))
axis(2, at = 0:25, las = 2)
```

```
hist(df2$perc_alcohol, breaks = seq(14.5, 49.5, 5),
     right = FALSE, xlab = "right open", font.lab = 2, axes = FALSE)
axis(1, at = seq(14.5, 49.5, 5))
axis(2, at = 0:25, las = 2)
```



My strategy is that since the biggest difference for the two plots is to include different boundary values, then setting the boundaries of the bars to be decimal values could let non of them include the boundary values thus making the two plots identical.

3.Titanic Survival

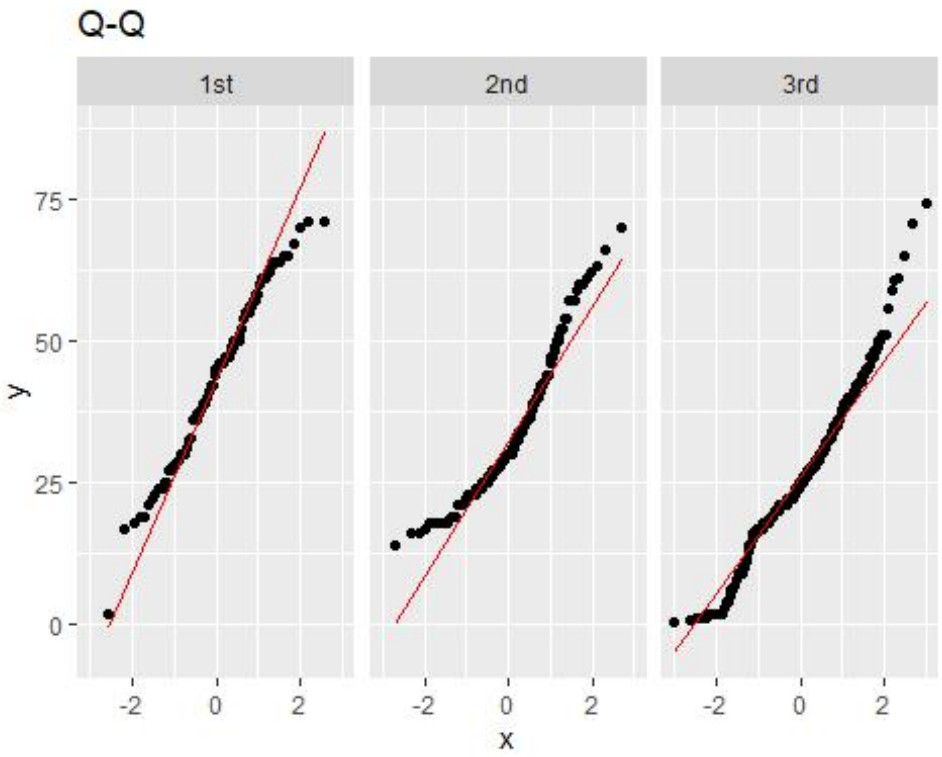
[8 points]

Data: *TitanicSurvival* in **carData** package

#a) Use QQ (quantile-quantile) plots with theoretical normal lines to compare age of **passengers who did not survive from Titanic** for the three different levels of `passengerClass`. What are some findings and for which class does the distribution of the age variable appear to be closest to a normal distribution?

```
#install.packages("perc_alcohol")
library(carData)
library(tibble)
library(ggplot2)
library(tidyverse)
df3<- tibble(TitanicSurvival)

df3 %>%
  filter(survived == "no") %>%
  ggplot(aes(sample=age)) +
    geom_qq(distribution = qnorm) +
    stat_qq_line(distribution = qnorm, color="red") + #, size = 1
    facet_wrap(~passengerClass) +
  labs (title = "Q-Q")
```

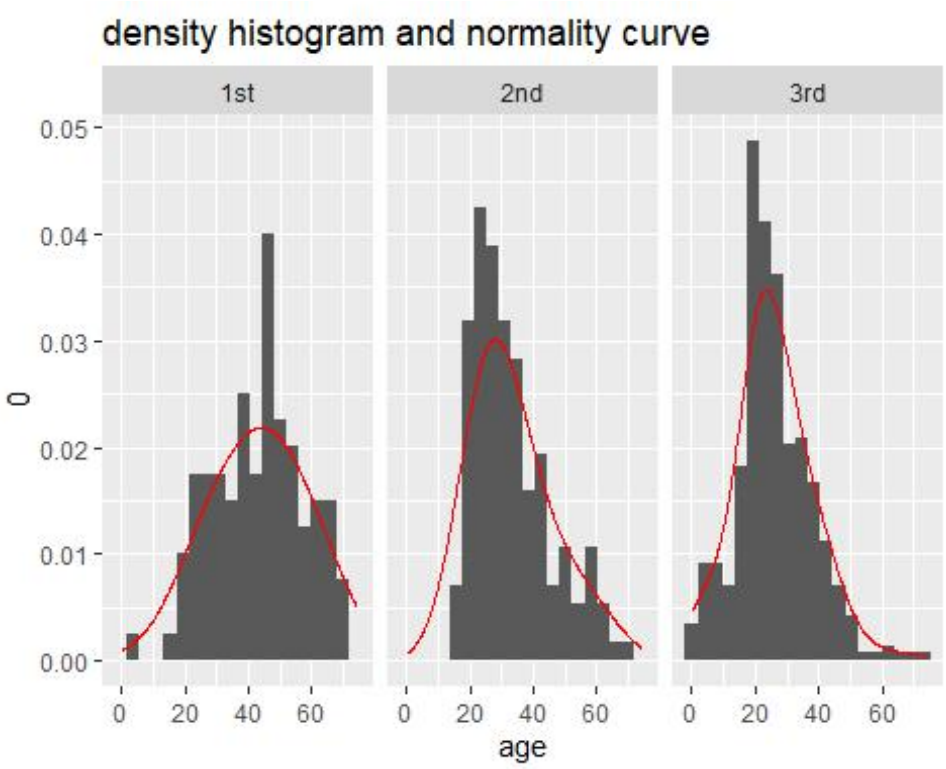
```
## Warning: Removed 190 rows containing non-finite values (stat_qq).

## Warning: Removed 190 rows containing non-finite values (stat_qq_line).
```



It can be seen from the three Q-Q plots that the age first class victims is more likely to be normally distributed, while the other two classes are not.

#b) Draw density histograms with density curves and theoretical normal curves overlaid of age for the three passenger classes.

```
df3 %>%
  filter( survived == "no") %>%
  ggplot(aes(x = age)) +
    geom_histogram(bins = 20, position="dodge", aes(y = ..density..)) +
    geom_density(colour="red", adjust = 2) +
    scale_y_continuous((limits = c(0, 0.04))) +
    facet_wrap(~passengerClass ) +
  labs( title = "density histogram and normality curve")

## Warning: Removed 190 rows containing non-finite values (stat_bin).

## Warning: Removed 190 rows containing non-finite values (stat_density).
```



Comparing to second and third class, the first class is still more likely to be normally distributed.

#c) Use a statistical method of your choice, such as the Shapiro-Wilk test, to determine which age distribution is closest to a normal distribution.

```
first <- df3 %>%
  filter( survived == "no") %>%
  filter( passengerClass == "1st")
f<-  shapiro.test(first$age)

second <- df3 %>%
  filter( survived == "no") %>%
  filter( passengerClass == "2nd")
s<- shapiro.test(second$age)

third <- df3 %>%
  filter( survived == "no") %>%
  filter( passengerClass == "3rd")
t<- shapiro.test(third$age)

f$p.value

## [1] 0.3173558

tibble("first class p"  = f$p.value,
       "second class p" = s$p.value,
       "third class p"  = t$p.value)

## # A tibble: 1 x 3
##   `first class p` `second class p` `third class p`
##             <dbl>            <dbl>           <dbl>
## 1           0.317      0.000000813      0.00000192
```

```
tibble("first class num"  = nrow(first),
       "second class num" = nrow(second),
       "third class num"  = nrow(third))
```

```
## # A tibble: 1 x 3
##   `first class num` `second class num` `third class num`
##               <int>              <int>             <int>
## 1               123                158               528
```

```
#
```

The p-value of the first class is bigger than 0.05, we accept the null hypothesis that the data is normally distributed. On the contrary, second and third class are not normally distributed as their p-value are much less than 0.05.

#d) Did all of the methods for testing for normality (a, b, and c) produce the same results? Briefly explain.

All the methods for testing give the same result that the age of first class victims is more likely to be normally distributed. This could due to the fact that the age of all passengers from first class, survived or not survived, is already normally distributed while passengers from other classes could concentrate more on younger people or middle aged people. Or to say, the younger and middle aged people in second and third class helped more older people to get out of the ship so that less old people died in the end, causing the age to be more like a positively skewed distribution.

4. Birds

[8 points]

Data: *birds* in **openintro** package

#a) Use appropriate techniques to describe the distribution of the `speed` variable noting interesting features.

```
#install.packages("openintro")
library(openintro)
```

```
## 载入需要的程辑包：airports
```

```
## 载入需要的程辑包：cherryblossom
```

```
## 载入需要的程辑包：usdata
```

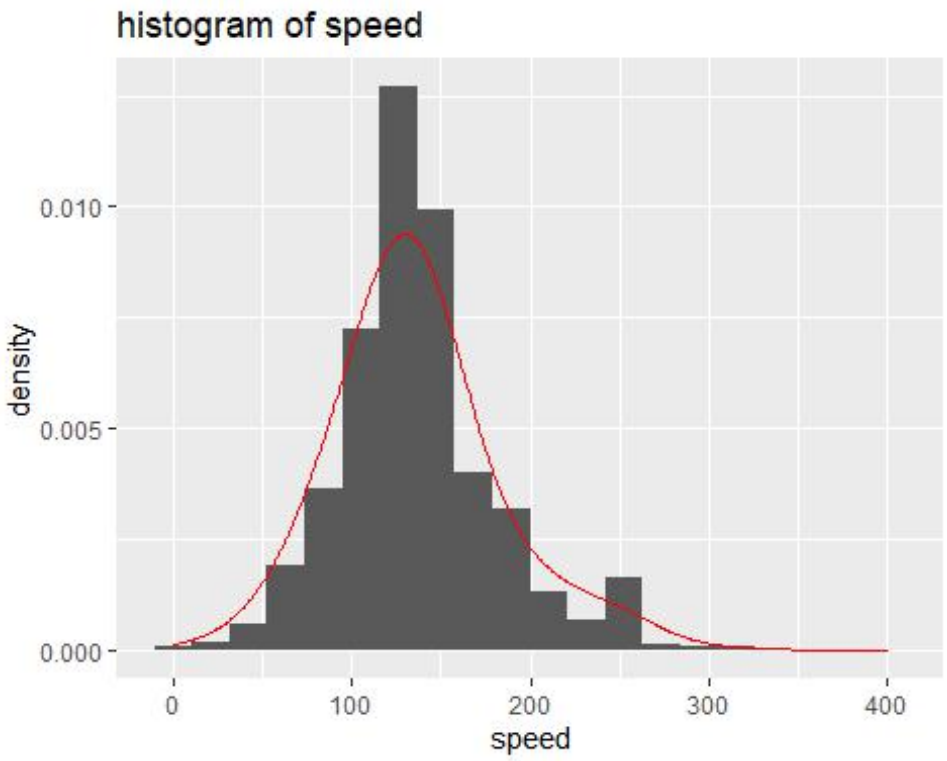
```
##
## 载入程辑包：'openintro'
```

```
## The following object is masked from 'package:fivethirtyeight':
##
##     drug_use
```

```
df4<- tibble(birds)
df4 %>% ggplot(aes(x = speed)) +
  geom_histogram(bins = 20, aes(y = ..density..)) +
  geom_density(colour="red", adjust = 6) +
  labs( title = "histogram of speed")
```

```
## Warning: Removed 7008 rows containing non-finite values (stat_bin).
```

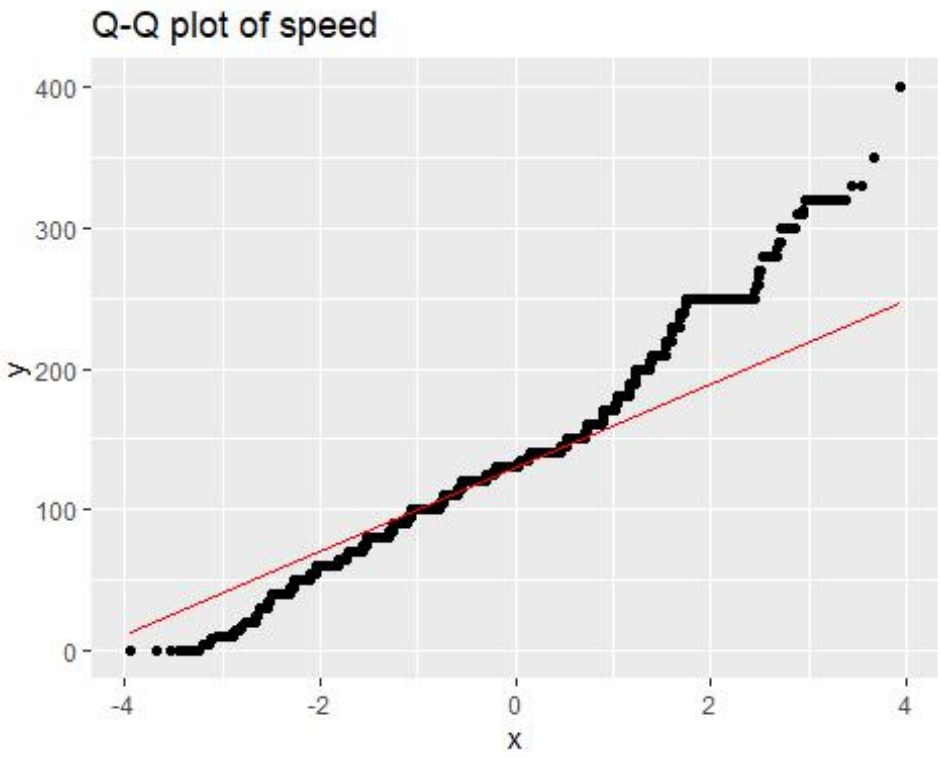
```
## Warning: Removed 7008 rows containing non-finite values (stat_density).
```




```
# looks like normally distributed
df4 %>%
  ggplot(aes(sample= speed)) +
  geom_qq(distribution = qnorm) +
  stat_qq_line(distribution = qnorm, color="red") + #, size = 1
  labs( title = "Q-Q plot of speed")
```

```
## Warning: Removed 7008 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 7008 rows containing non-finite values (stat_qq_line).
```




```
ks.test(df4$speed, "pnorm")
```

```
## Warning in ks.test(df4$speed, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df4$speed
## D = 0.99935, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

From the density histogram, we can see that the most airplanes have a speed of 110 to 150, and planes with a speed of 125 (approximately) are the most. Also, there is a little gap between 210 and 250, with a lot less planes having a speed of 240 (approximately) but a lot more planes with a speed of 250.

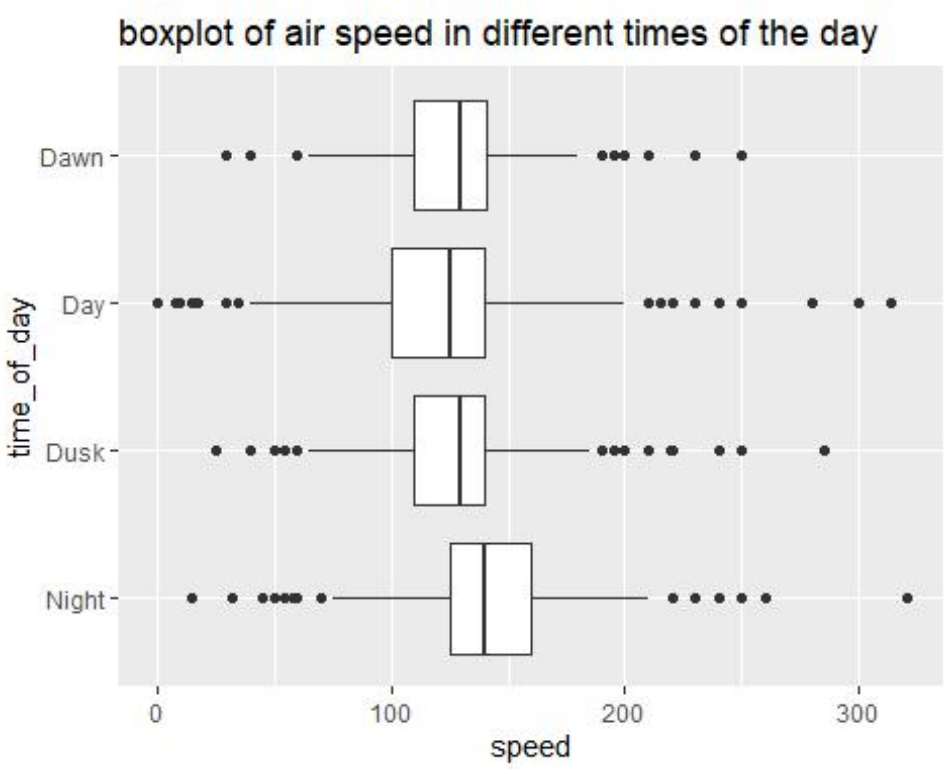
It seems like a normal distribution from the density histogram, but from Q-Q plot and KS-test we know the data is not normally distributed.

#b) Create horizontal boxplots of `speed`, one for each level of `time_of_day`.

```
unique(df4$time_of_day)
```

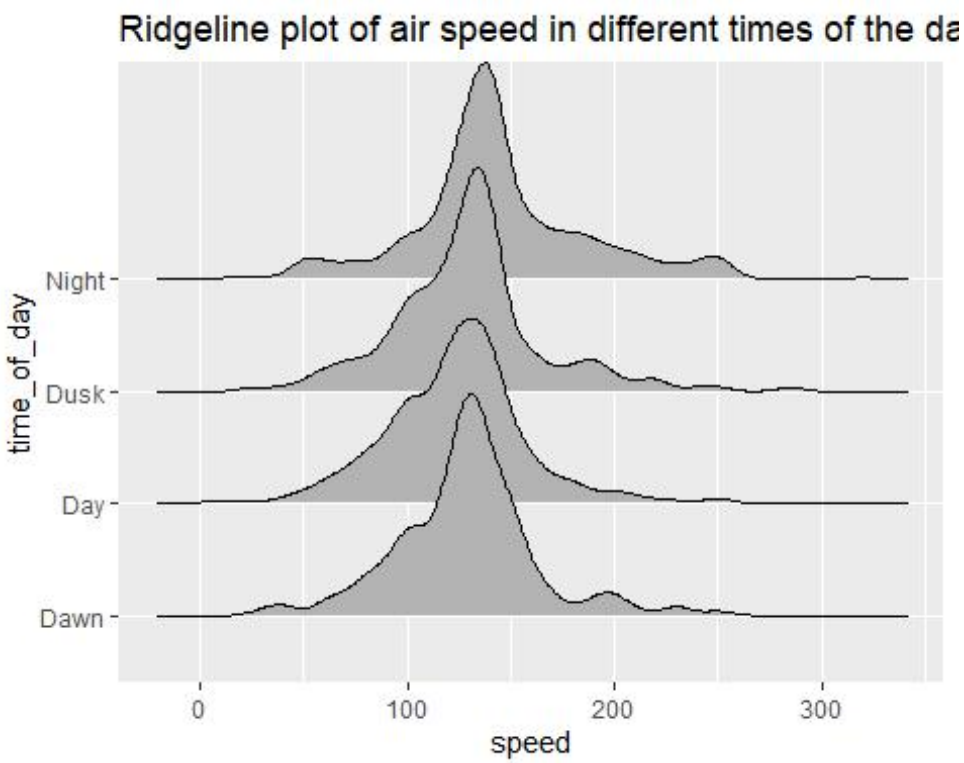
```
## [1] Night Day   Dusk  Dawn  <NA>
## Levels: Dawn Day Dusk Night
```

```
df4 %>%
  mutate( time_of_day = fct_relevel( time_of_day, "Night",  "Dusk", "Day", "Dawn")) %>%
  na.omit() %>%
  ggplot (aes(time_of_day, speed))+
  geom_boxplot() +
  coord_flip() +
  labs (title = "boxplot of air speed in different times of the day")
```



#c) Create ridgeline plots for the same data as in b)

```
#install.packages("remotes")
library(remotes)

#remotes::install_github("R-CoderDotCom/ridgeline@main")

library(ridgeline)

#install.packages("ggridges")
#remotes::install_github("wilkelab/ggridges")

library(ggridges)
df4 %>%
  na.omit() %>%
  ggplot(aes(x =speed, y = time_of_day)) +
  geom_density_ridges(scale = 2) +
  labs( title = "Ridgeline plot of air speed in different times of the day"  )

## Picking joint bandwidth of 7.01
```



#d) Compare the boxplot plots and the ridgeline plots.

The boxplot can give us a clear view of how the median value of speed are different at different times of the day and also tell us how many outliers are there, which cannot be shown by the ridgeline plot. However, the ridgeline plot can tell us the distribution of the different times and how different their mean values are, which cannot be told by the boxplot.