Case study – The Titanic disaster CMIBOD021T – G. Costantini

Important notes:

- ✓ You can work in groups of *two* students, but the final report is *individual*. The R *code* and the *plots/visualization* can be the same, but the *comments/explanations/answers* must be different from each other.
- ✓ After the delivery, the teacher has the possibility to ask you to explain (in person) what you wrote in this report, to check the autenticity of the contents.
- ✓ Fill this document with your answers (code, outputs, explanations, etc…) and upload it on N@tschool *before* the deadline: midnight of Sunday 08 November 2015 (end of week 9).

Student name: Jeroen van der Steen

Student number: 0867254

School/Opleiding: Hogeschool Rotterdam, Media Technologie

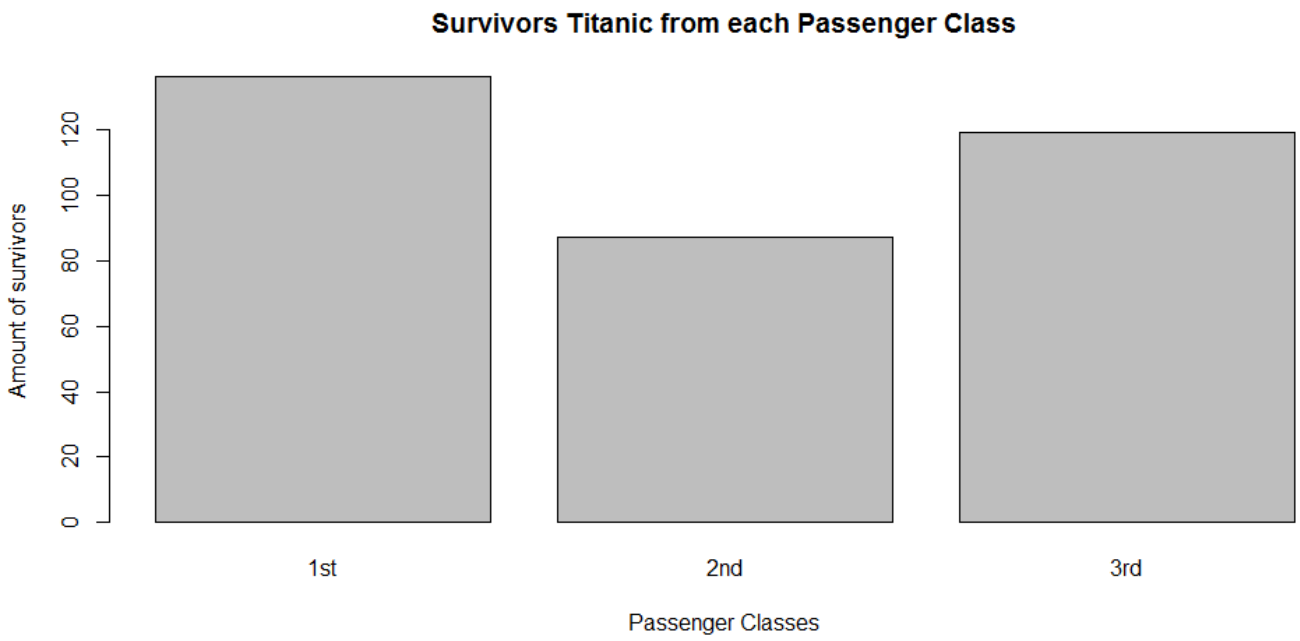
a) [*Week 1; Chapters 1,2 Dalgaard*] Download the dataset "train.csv" from the Kaggle website (https://www.kaggle.com/c/titanic-gettingStarted/data) or from N@tschool. Read from the website the columns description. Import the dataset into R and examine its contents.

```
VARIABLE DESCRIPTIONS:

survival        Survival

                (0 = No; 1 = Yes)

pclass          Passenger Class

                (1 = 1st; 2 = 2nd; 3 = 3rd)

name            Name

sex             Sex

age             Age

sibsp           Number of Siblings/Spouses Aboard

parch           Number of Parents/Children Aboard

ticket          Ticket Number

fare            Passenger Fare

cabin           Cabin

embarked        Port of Embarkation

                (C = Cherbourg; Q = Queenstown; S = Southampton)



SPECIAL NOTES:

Pclass is a proxy for socio-economic status (SES)
```

```
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower


Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5


With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.


Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic


Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

b) [24p] [*Week 2; Chapter 4 Dalgaard*] Create some visualizations (boxplots, histograms, barplots, tables, etc…) to represent *interesting* information about the dataset. For example, you could produce a barplot (remember the option "beside") showing the proportion of people which survived with respect to their passenger class. Produce at least 4 visualizations. For each visualization, explain the meaning of the values (i.e., what do the rows/columns/bars/… values represent) and what information you can infer from the plot/table.

[6p] Visualization 1:

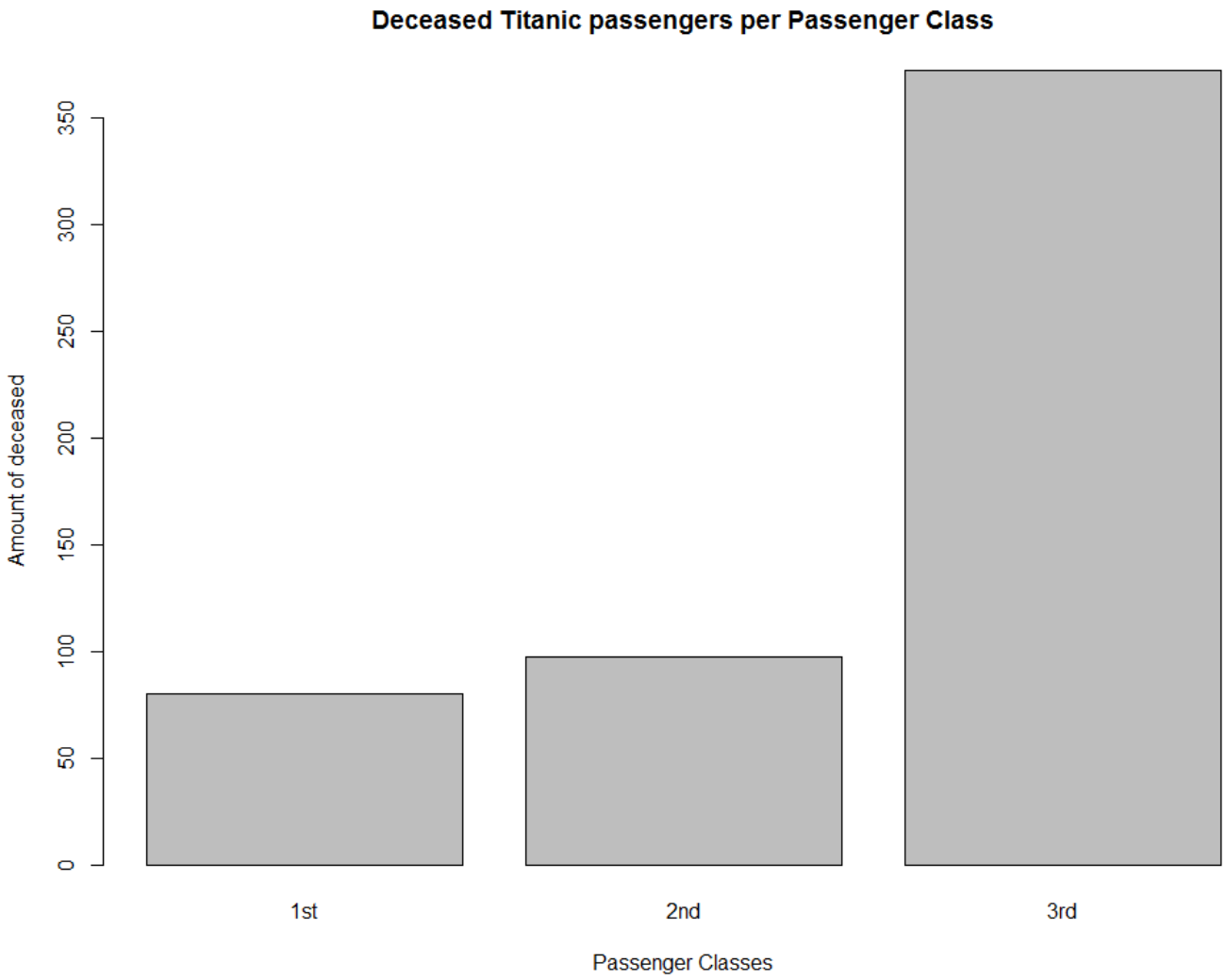



```
#Find all survivors; 342
survivors <- train[train$Survived==1,]
#Find the Passenger Classes; 1,2,3
levels(factor(survivors$Pclass))

#b1: survivors per passanger class
barplot(
  height = c(
    sum(survivors$Pclass==1),
    sum(survivors$Pclass==2),
    sum(survivors$Pclass==3)
  ),
  names.arg=c(
    "1st","2nd","3rd"
  ),
  xlab="Passenger Classes",ylab="Amount of survivors",
  main="Survivors Titanic from each Passenger Class"
)
```

Comment:

The Barplot shows the amount of survivors from the Titanic, from each Passenger class.
The first class had the most survivors.

[6p] Visualization 2:

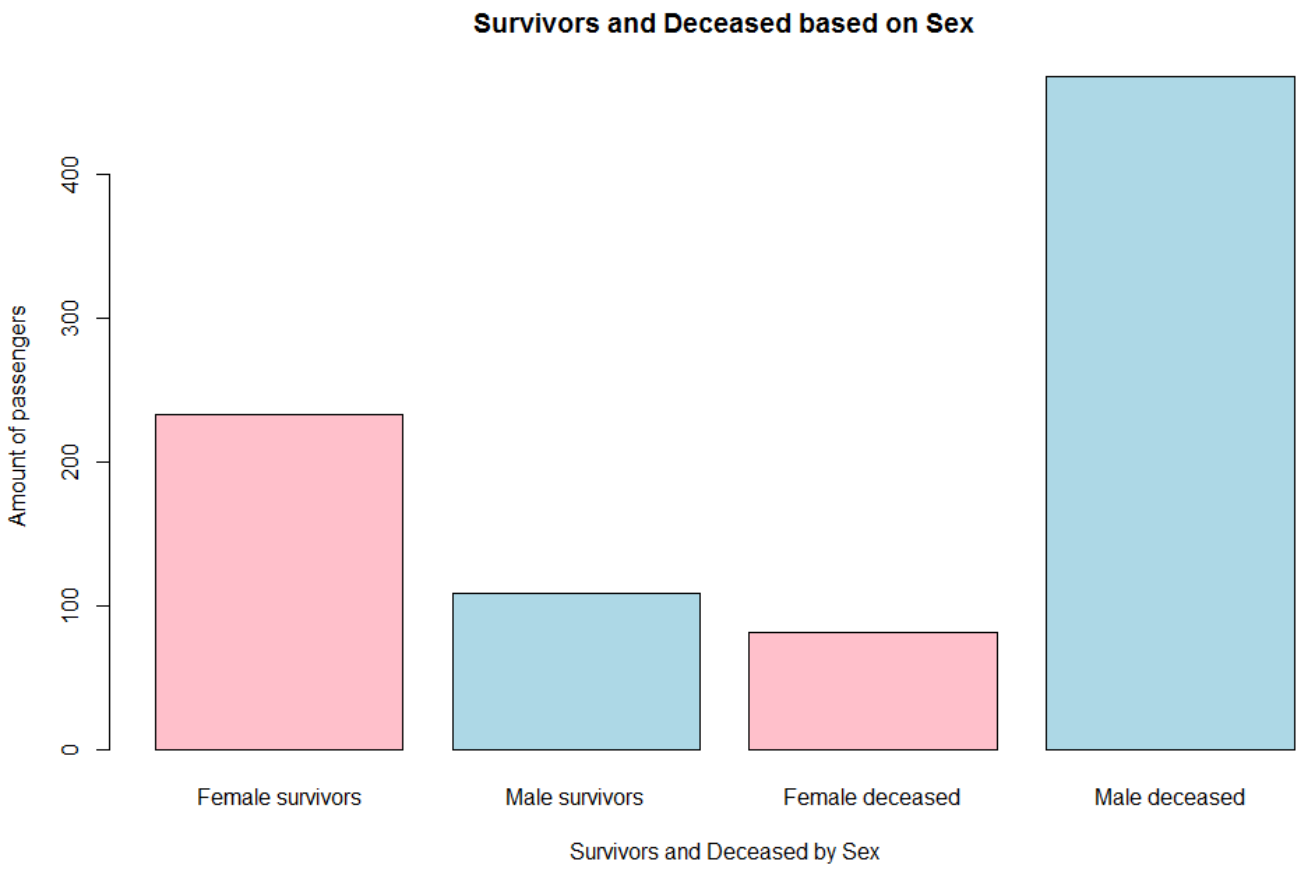



```
#Find all deceased; 549
deceased <- train[train$Survived==0,]
#b2: deceased per passanger class
barplot(
  height = c(
    sum(deceased$Pclass==1),
    sum(deceased$Pclass==2),
    sum(deceased$Pclass==3)
  ),
  names.arg=c(
    "1st","2nd","3rd"
  ),
  xlab="Passenger Classes",ylab="Amount of deceased",
  main="Deceased Titanic passengers per Passenger Class"
)
```

Comment:

The Barplot shows the amount of deceased passengers from the Titanic, per each Passenger class.
The third class had the highest death toll.

[6p] Visualization 3:

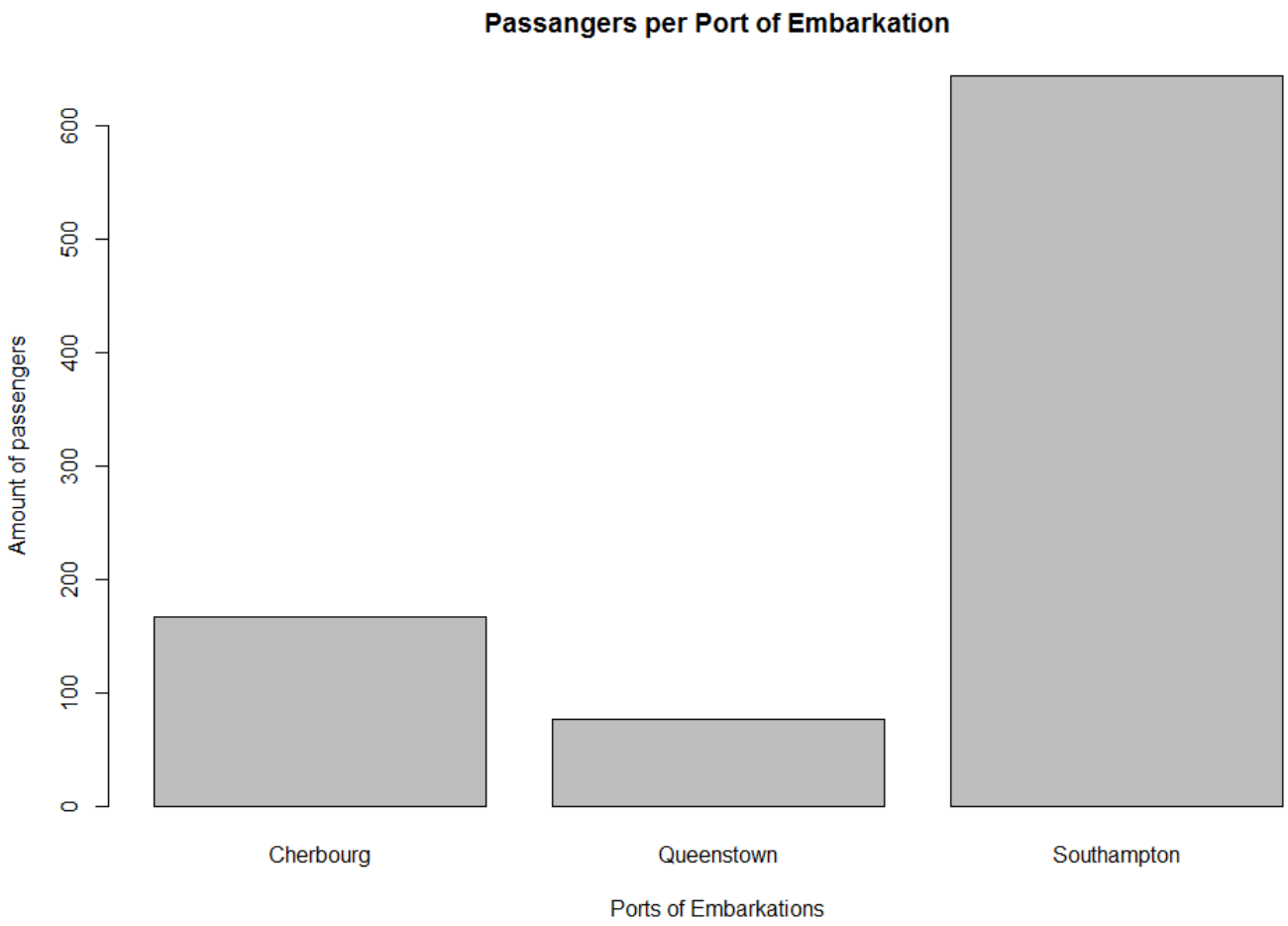



```
female_survivors <- survivors[survivors$Sex == "female",]
male_survivors <- survivors[survivors$Sex == "male",]
female_deceased <- deceased[deceased$Sex == "female",]
male_deceased <- deceased[deceased$Sex == "male",]

library(plyr)
female_survivors <- count(female_survivors, c("Survived"))$freq
male_survivors <- count(male_survivors, c("Survived"))$freq
female_deceased <- count(female_deceased, c("Survived"))$freq
male_deceased <- count(male_deceased, c("Survived"))$freq

barplot(
  c(female_survivors, male_survivors, female_deceased, male_deceased),
  names.arg=c("Female survivors", "Male survivors", "Female deceased", "Male deceased"),
  main="Survivors and Deceased based on Sex",
  xlab="Survivors and Deceased by Sex",
  ylab="Amount of passengers",
  col=c("pink", "lightblue", "pink", "lightblue")
)
```

Comment: *[explanation about the visualization]*

The Barplot shows the amount of survivors and deceased per gender.
The first two bars show the survivors, the 3th and 4th bar show the deceased passangers.
Both genders have a corresponding colour.
You can see that more male passangers deceased then female, and that more female passangers survived then male.

[6p] Visualization 4:




```
embarkations <- train[train$Embarked != "",]
embarkations <- count(embarkations$Embarked)
barplot(embarkations$freq,
    names.arg=c("Cherbourg", "Queenstown", "Southampton"),
    main="Passangers per Port of Embarkation",
    xlab="Ports of Embarkations",
    ylab="Amount of passengers"
    )
```

Comment: *[explanation about the visualization]*

The visual show the amount of passangers per port of embarkation.
The most passangers entered the ship from the Southampton port.

c) [12p] [*Week 4; Chapters 3,4 Dalgaard*] Consider the Age variable. Compute its mean and its standard deviation. Then, put in the same plot:

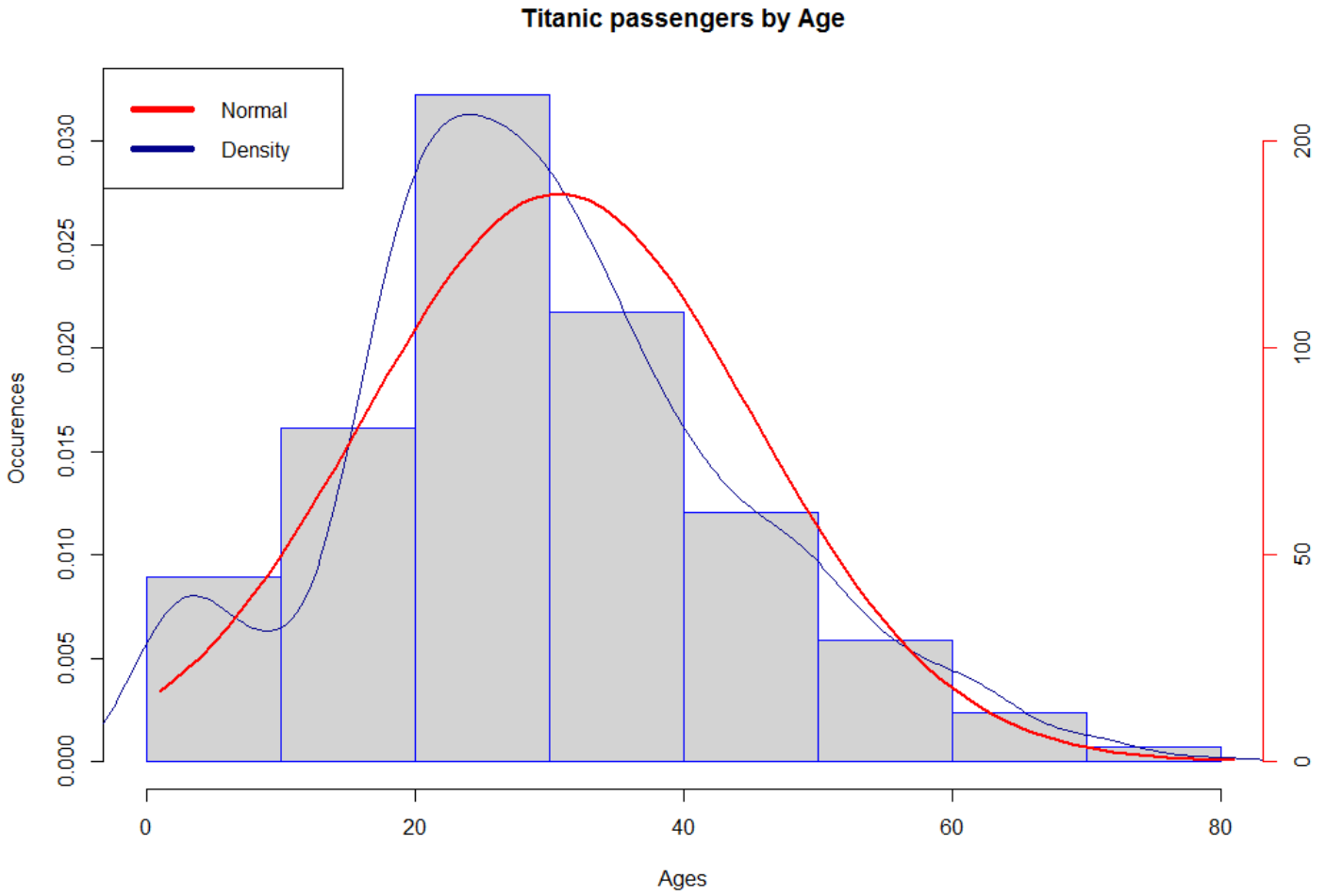
- the density function of the Normal distribution, using as parameters the mean and standard deviation of Age;
- a histogram of the Age variable.

For the histogram, remember that you can specify the starting position of the bars using the option "*breaks*". To add a histogram to an existing plot, use the option "*add=T*". Try to make the plot as good-looking as possible.
Then, produce also the Q-Q plot for the Age variable.

[1p] Mean of Age: 29.69912

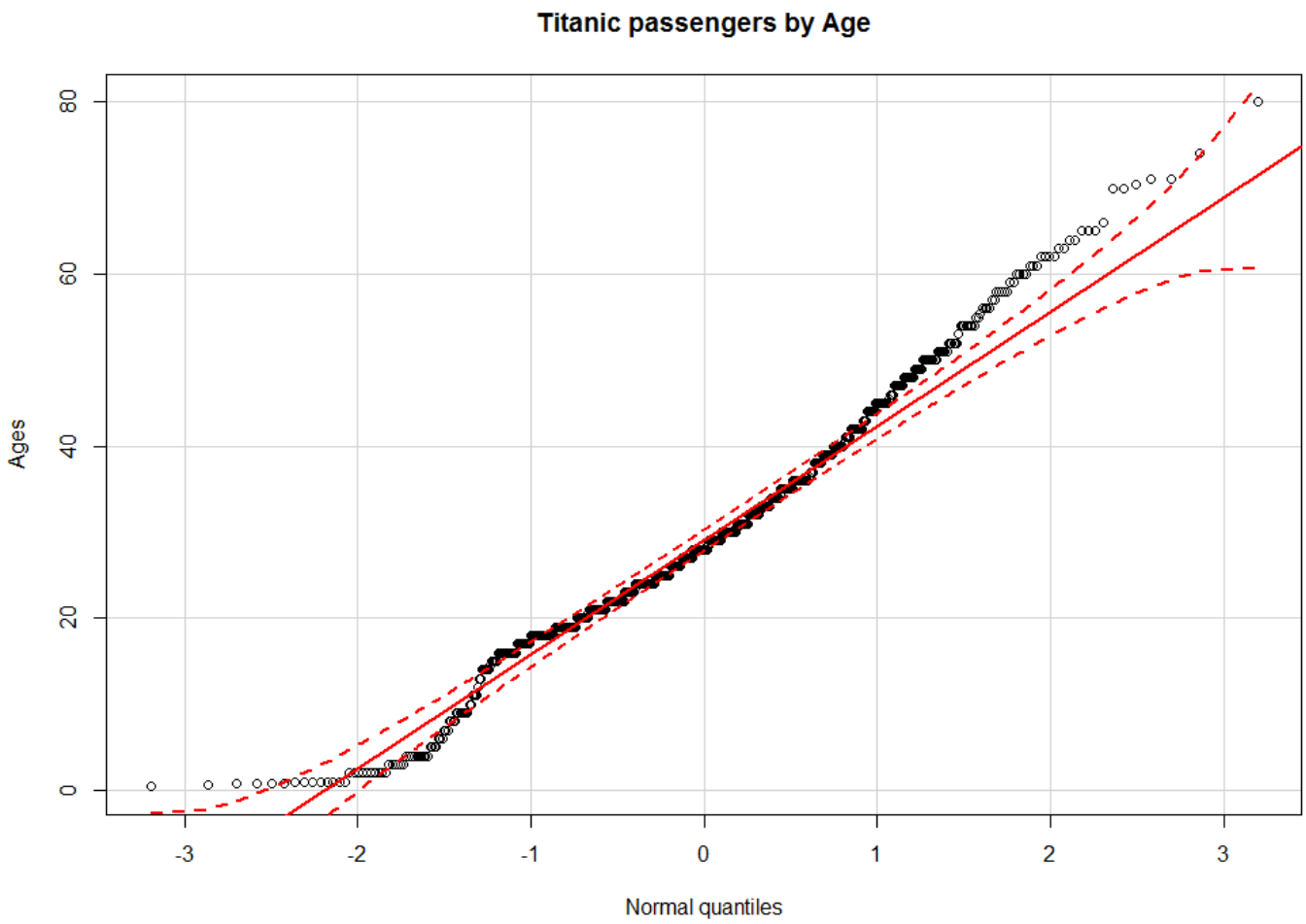
[1p] Standard deviation of Age: 14.5265

[5p] Plot of Normal distribution and histogram of Age (together in only one plot!):




```
hist(train$Age, freq=F, col="lightgrey", xlab="Ages", ylab="Occurences", main="Titanic passengers by Age", border="blue")
lines(density(train$Age, na.rm=T), col="darkblue")
lines(dnorm(0:80,mAge,sdAge), col="red", lwd=2)
legend("topleft", c("Normal", "Density"), col=c("red", "darkblue"), lwd=5)
axis(side = 4, col = "red", at=c(0,.01,.02,.03), labels=c(0,50,100,200))
```

[2p] Q-Q plot for the Age variable:



qqnorm(train$Age, datax=T, ylab = "Years", xlab="Occurences", main="Titanic passengers by Age")

install.packages("car")

library(car)

qqPlot(train$Age, main="Titanic passengers by Age", ylab="Ages", xlab="Normal quantiles")

[3p] Looking at the two plots, do you think that the Age variable is normally distributed? Why?

No, you can see a peak between the age of 20 and 30 years old. This age group is kind of in the center, **but** it's not a complete bell shape. There are more passangers older than 40 years old, then people younger than 20 years old.

d) [9p] [*Week 4; Chapter 6 Yakir*] Using the results obtained at point c), suppose that Age is really distributed as a Normal random variable and compute:

[3p] The probability that a passenger is younger than 12

→ R code:

mAge <- mean(train$Age, na.rm=T)

sdAge <- sd(train$Age, na.rm=T)

pnorm(12,mAge,sdAge)

→ Result: 0.1115356

[3p] The probability that a passenger is between 20 and 50 years old

→ R code:
pnorm(50,mAge,sdAge) - pnorm(20,mAge,sdAge)

→ Result: 0.6667019

[3p] The probability that a passenger is older than 65

→ R code:
1 - pnorm(65,mAge,sdAge)

→ Result: 0.00754727

e) [40p] [*Week 6; Chapter 5 Dalgaard*] Execute statistical tests to answer the following research questions. [Note: in order to execute some of the tests, you could need to create new vectors of values starting from the original dataset variables]

[10p] Research question 1: Was there a different **survival rate** between men and women? In other words, did you have more chances of being saved depending on your sex?

Yes, you have more chance of being saved depending on your sex.

Females are more likely to be saved.

females <- train[train$Sex == "female",]

males <- train[train$Sex == "male",]

t.test(females$Survived, males$Survived, alternative="two.sided")

Answer: *[motivate the answer based on values from the R output]*

We reject the null hypothesis (that gender doesn't matter if you survive or not), because 0 is not included in the confidence interval (0.4949481 and 0.6113121).

[15p] Research question 2: Was there a different survival rate between the "critical age" people and the other? Let us define the "critical age" as being a child (≤ 8 years old) or a senior (≥ 65 years old). In other words, did you have more chances of being saved if you were in a "critical age"?

No, you don't have more chance of being saved if you were in the critical age group.

criticals <- train[train$Age <= 8 | train$Age >= 65,]

noncriticals <- train[train$Age >= 8 | train$Age <= 65,]

t.test(noncriticals$Survived, criticals$Survived, alternative="greater")

Answer: *[motivate the answer based on values from the R output]*

The 0 is included in the confidence interval (-0.2705962 and Inf), so we accept the null hypothesis.

[15p] Research question 3: Was there a different survival rate between the people of different passenger classes? In other words, did you have more chances of being saved if you were in an upper class? [Careful: there are 3 passenger classes but you cannot make a test on 3 vectors at the same time...]

You can say yes, if you are in a higher class you have a better chance of that you will survive.

ps1 <- train[train$Pclass == 1,]

ps2 <- train[train$Pclass == 2,]

ps3 <- train[train$Pclass == 3,]

sum(ps3$Survived == T) #119

sum(ps2$Survived == T) #87

sum(ps1$Survived == T) #136

#Alternative: Is a/Higher class survivors greather then b/Opposite?

t.test(ps1$Survived, ps2$Survived, alternative="greater") #Rejected null hypothesis, 0 is not in conf

t.test(ps1$Survived, ps3$Survived, alternative="greater") #Rejected null hypothesis, 0 is not in conf

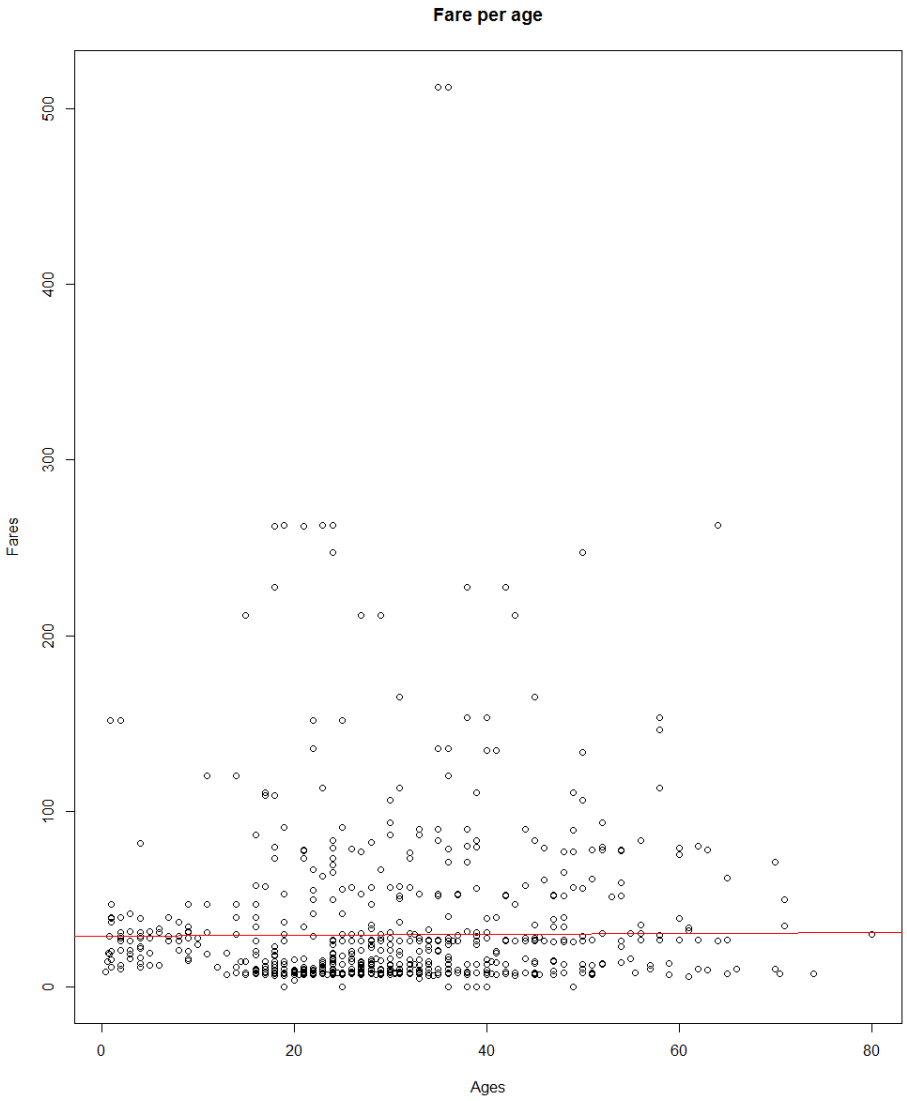
t.test(ps2$Survived, ps1$Survived, alternative="greater") #Accepted null hypothesis, 0 is in conf

Answer: *[motivate the answer based on values from the R output]*

The higher class only haves a better chance of survival two times, you can see this because the 0 is only onces in the confidence range. And two times the 0 is not in the confidence range.

f) [15p] [*Week 7; Chapter 6 Dalgaard*] Consider the variables Fare and Age. Plot the fare in function of the age. Fit a linear regression model on these two variables and add the resulting line on the plot. Get the summary of the model and investigate on the existence of a relationship between the two variables.

[5p] Linear regression:



[2p] Summary of the model:

```
Call:
lm(formula = ages ~ fares)

Residuals:
    Min      1Q  Median      3Q     Max
-31.861  -9.024  -1.134   8.019  50.425

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.78420    0.64765  44.444   <2e-16 ***
fares        0.02637    0.01024   2.575   0.0102 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.47 on 712 degrees of freedom
  (177 observations deleted due to missingness)
Multiple R-squared:  0.009229, Adjusted R-squared:  0.007837
F-statistic: 6.632 on 1 and 712 DF,  p-value: 0.01022
```

[1p] Value of the intercept of the model:	28.78420
[1p] Value of the slope of the model:	0.02637
[1p] Equation of the linear regression model:	28.78420 + 0.02637 * fare

[5p] From the resulting model summary (and looking at the plot), do you think that there is a relationship between Fare and Age? In other words, did you pay consistently less (or more) for your ticket, depending on your age? [*motivate your answer referring to specific numbers/features of the summary/plot*]

I think you have more people that are between 20 and 60 years old, this is what the plot is showing.
But you don't really pay more then. The abline-line also doesn't seem to be right, it could be more precise if it was curved.