报告题目：**Final Exam Project**

姓　　名：陈 睿 15620161152244

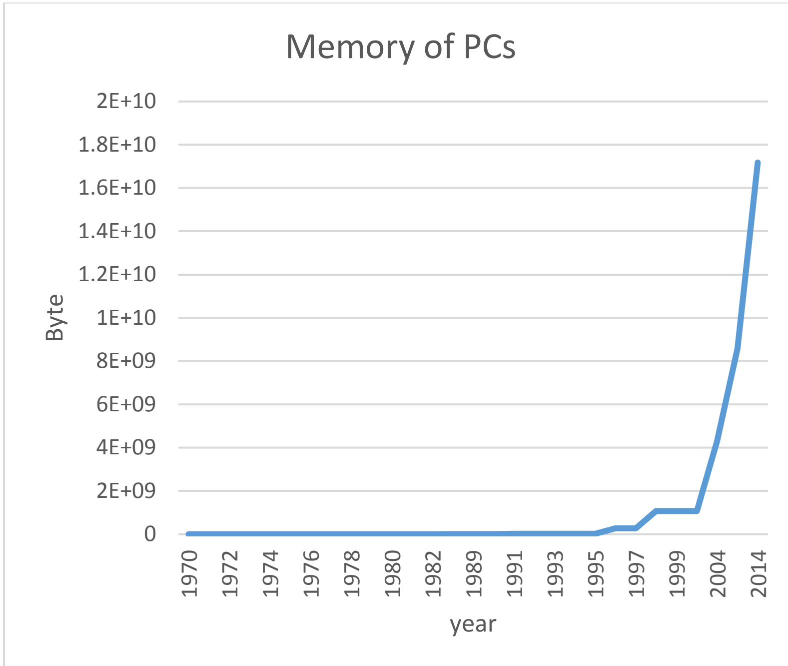
任课老师：Wolfgang 教授

日　　期：2017 年 10月 28 日

HW Unit1

1.1 Calculate the increase of memory of PCs over the last 30 years and check whether the FMRI analysis could have been done 20 years ago.

year	**Byte**
1970	262144
1971	262144
1972	262144
1973	262144
1974	262144
1975	262144
1976	262144
1977	262144
1978	262144
1979	262144
1980	262144
1981	262144
1982	262144
1988	2097152
1989	2097152
1990	2097152
1991	16777216
1992	16777216
1993	16777216
1994	16777216
1995	16777216
1996	268435456
1997	268435456
1998	1073741824
1999	1073741824
2000	1073741824
2004	4294967296
2009	8589934592
2014	17179869184



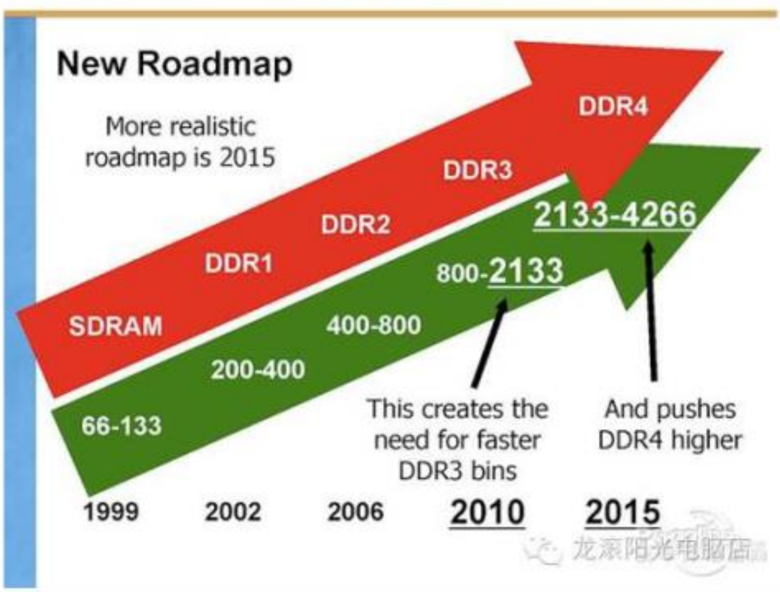
Comments: There was the rather slow growth before 2000, while there is dramatically great increase after 2000.

1.2 Prepare 2-5 slides explaining logistic regression.

Logistics regression model

Rui CHEN

The development of home computer inter storage



Generalized Linear Models

- First of all, we briefly review the concept of generalized linear models (GLMs). Logistic regression is just one example of this type of model.

- All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable;
2. A linear model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$;
3. A link function that relates the linear model to the parameter of the outcome distribution:

$$g(p) = y \; or \; p = g^{-1}(y)$$

Logistic Regression

- Logistic regression is a GLM used to model a binary categorical variable using numerical and predictors.

- We assume a binomial distribution produced the outcome variable and we therefore want to model *p* the probability of success for a given set of predictors.

- To finish specifying the logistic model we just need to establish a reasonable link function that connects *y* to *p*. There are a variety of options but the most commonly used is the logit function.

$$logit(p) = log\left(\frac{p}{1-p}\right), \qquad for\; 0 \leq p \leq 1$$

Properties of the Logit

- The logit function takes a value between 0 and 1, mapping it to a value between $-\infty$ and ∞.

- Inverse logit (logistic) function

$$g^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

- The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

- This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds for a success.

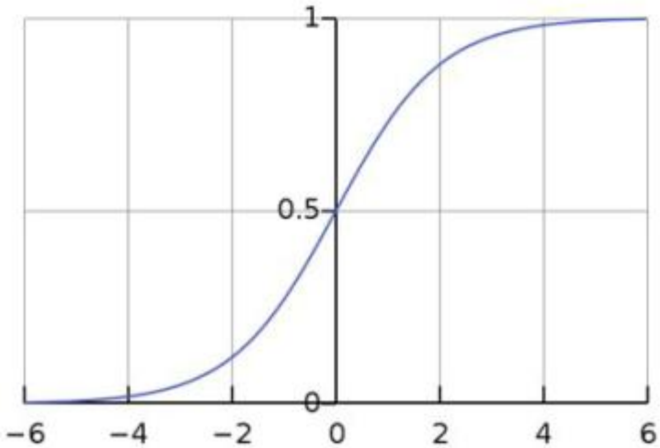
The logistic regression model

- The three GLM characteristics give us:
-
$$1.\ y_i \sim Binom(p_i)$$
$$2.\ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$
$$3.\ logit(p) = y$$

- From which we arrive at,

$$p_i = P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,t} + \cdots + \beta_n x_{n,t})}},$$
$$where\ \boldsymbol{x} = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})$$

Graph



1.3 Apply for an account for Github.
My account is 22457114@qq.com, RuiChan244.

HW Unit 2

2.1 Make an R Quantlet to solve HW #1 from unit 1 with R and show it on Github (GH). Hint: use the CMB Qs for this work.

memory.df = read.csv("byte.csv",header = TRUE)

plot(memory.df$Byte~memory.df$year,type="o",main="The development of internal memory")

2.2 Use R with B-spline code to solve HW#1, any comments?

splines.reg.l1 = smooth.spline(x = memory.df$year, y = memory.df$Byte, spar = 0.2)

splines.reg.l2 = smooth.spline(x = memory.df$year, y = memory.df$Byte, spar = 1)

splines.reg.l3= smooth.spline(x = memory.df$year, y = memory.df$Byte, spar = 2)

lines(splines.reg.l1, col = "green", lwd = 2)

lines(splines.reg.l2, col = "pink", lwd = 2)

lines(splines.reg.l3, col = "blue", lwd = 2)

2.3 Suppose you observe that in n=1000 mails (in 1 week) you have about 2 scams. Use the LvB /Poisson cdf to calculate that you have 6 scam emails in 2 weeks. In Scammyland you have 5 scams on average, what is the probability to have no scam mail.

lambda=2

x=3

probex1=exp(-lambda)*lambda^x/factorial(x)

probex1

```
> lambda=2
> x=3
> probex1=exp(-lambda)*lambda^x/factorial(x)
> probex1
[1] 0.180447
```

lambda=5

x=0

probex2=exp(-lambda)*lambda^x/factorial(x)

probex2

```
> lambda=5
> x=0
> probex2=exp(-lambda)*lambda^x/factorial(x)
> probex2
[1] 0.006737947
```

HW Unit 3

3.1 Make an R quantlet on GH to produce hash code for the 2 sentences: „I learn a lot from this class when I am proper listening to the professor“, I do not learn a lot from this class when I am absent and playing on my Iphone“. Compare the 2 hash sequences.

install stuff for hash calculation

install.packages("digest")

call the library doing the hashes

library("digest")

digest("I learn a lot from this class when I am proper listening to the professor")

#"a8d3e4701672195e5dcd16ea9b062279"

digest("I do not learn a lot from this class when I am absent and playing on my phone")

#"059ab10d478614d2eab3d70cfccd3fcc"

digest("I learn a lot from this class when I am proper listening to the professor","sha256")

#"c16700de5a5c1961e279135f2be7dcf9c187cb6b21ac8032308c715e1ce9964c"

digest("I do not learn a lot from this class when I am absent and playing on my phone","sha256")

#"f5e2cba48dac097355d0bb310fdbd5bd38a22a5c8e8215cd1ae67014cfc35b91"

3.2 Make 3-5 slides (in PPTX) on the DSA (Digital Signature Algorithms)

DSA
(Digital Signature Algorithms)

Presented by Rui Chen 15620161152244

Department of Finance, SOE

Definition of DSA

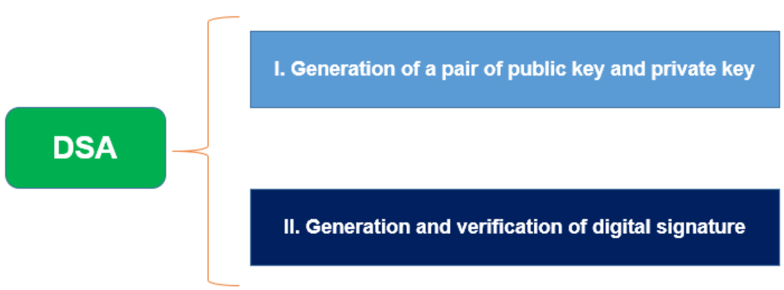
The Digital Signature Algorithm (DSA) is a Federal Information Processing Standard for digital signatures. In August 1991 the National Institute of Standards and Technology (NIST) proposed DSA for use in their Digital Signature Standard (DSS) and adopted it as FIPS 186 in 1993. Four revisions to the initial specification have been released: FIPS 186-1 in 1996, FIPS 186-2 in 2000, FIPS 186-3 in 2009, and FIPS 186-4 in 2013.

Definition of DSA

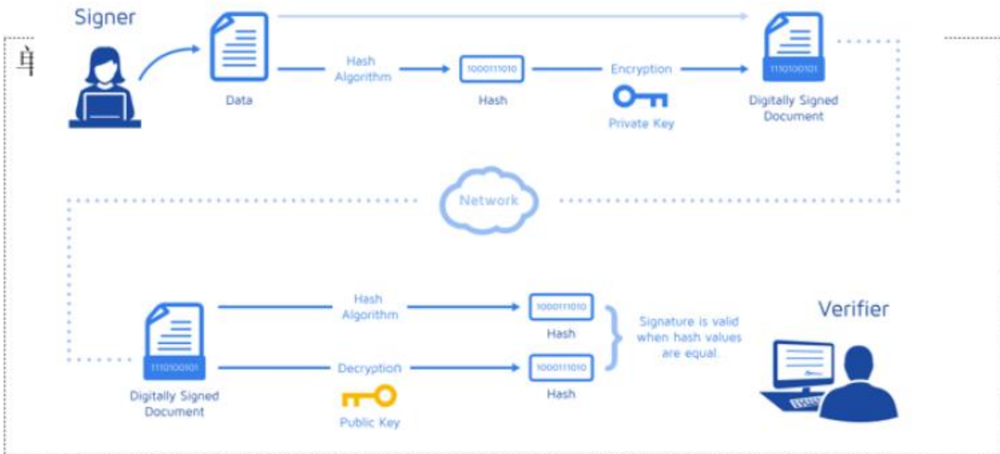
DSA is covered by U.S. Patent 5,231,668, filed July 26, 1991 and attributed to David W. Kravitz, a former NSA employee. This patent was given to "The United States of America as represented by the Secretary of Commerce, Washington, D.C.", and NIST has made this patent available worldwide royalty-free. Claus P. Schnorr claims that his U.S. Patent 4,995,082 (expired) covered DSA; this claim is disputed. DSA is a variant of the ElGamal signature scheme.

——From Wikipedia

DSA consists of the following two parts：



How do digital signatures work?



Key generation has two phases：

Two phases

- **I. A choice of algorithm parameters which may be shared between different users of the system**
- **II. Computing public and private keys for a single user**

The steps of performing the digital signature

1. **Calculate the Message Digest** (hash-value of the message)
In the first step of the process, a hash-value of the message (often called the message digest) is calculated by applying some cryptographic hashing algorithm.
2. Calculate the Digital Signature
In the second step of digitally signing a message, the information obtained in the first step hash-value of the message (the message digest) is encrypted with the private key of the person who signs the message and thus an encrypted hash-value, also called digital signature, is obtained. For this purpose, some mathematical cryptographic encrypting algorithm for calculating digital signatures from given message digest is used, which includes **DSA, TSA, ECDSA and so on**.
3. **Verifying Digital Signatures**
The public key is used in the signature verification process to verify the authenticity of the signature.

Reference：

https://en.wikipedia.org/wiki/Digital_Signature_Algorithm

Thank you!

3.3 Make slides with R code where you create a JSON data set that you save and read again.

```
install.packages("rjson", repos="http://cran.us.r-project.org")

library(rjson)

json_file = "http://crix.hu-berlin.de/data/crix.json" json_data = fromJSON(file=json_file)

x = as.data.frame(json_data) date1=c(json_data[[1]]$date)

for (i in 1:50){date1[i]=c(json_data[[i]]$date)}

price1=c(json_data[[1]]$price)

for (i in 1:50){ price1[i]=c(json_data[[i]]$price)}

date=date1

price=price1

crix=data.frame(date,price)

plot(crix$price~as.Date(crix$date))

plot(crix$price~crix$date,type="b")

plot(ts(crix$price,freq=1),type='l',xlab='Day',ylab='Price')
```

3.4 Download the CRIX data and make a plot of the time series, analyse its properties, i.e. fit ARMA, ARIMA etc. Is there a GARCH effect?

```
install.packages("caschrono ", repos="http://cran.us.r-project.org")

install.packages("TTR ", repos="http://cran.us.r-project.org")

install.packages("fGarch ", repos="http://cran.us.r-project.org")

install.packages("rugarch ", repos="http://cran.us.r-project.org")
```

```
install.packages("forecast ", repos="http://cran.us.r-project.org")

install.packages("TSA ", repos="http://cran.us.r-project.org")


library(caschrono)

library(TTR)

library(fGarch)

library(rugarch)

library(forecast)

library(TSA)

#******************ARIMA model*****************

xy.acfb(crix$price,numer=FALSE)

adf.test(crix$price)

#Augmented Dickey-Fuller Test:not stationary

##*****1)return

r=diff(log(crix$price))*100

plot(r,type="b")

abline(h = 0)

plot(r,type="l")

xy.acfb(r,numer=FALSE)

#*****2)Parameter Estimation

#estimation of p and q

a.fin2=arima(r,order=c(2,0,2))

summary(a.fin2) f=forecast(a.fin2,h=3,level=c(99.5)) acf(f$residuals,lag.max = 20)
Box.test(f$residuals,lag=20,type='Ljung-Box') #the residuals follow Gaussian distribution
plot.ts(f$residuals)

#****3)some evidence to GARCH model

#get ACF and PACF of the residuals xy.acfb(residuals(a.fin2),numer=FALSE)
xy.acfb((residuals(a.fin2))^2,numer=FALSE)+ xy.acfb(abs(residuals(a.fin2)),numer=FALSE)

#get the Conditional heteroskedasticity test McLeod.Li.test(y=residuals(a.fin2))

#p-values are all included in the test, it formally shows strong evidence for ARCH in this data.

#**Normality of the Residuals
```

qqnorm(residuals(a.fin2))

qqline(residuals(a.fin2))

shapiro.test(residuals(a.fin2))

#The QQ plot suggest that the distribution of returns may have a tail thicker that of a

#normal distribution and maybe somewhat skewed to the right

#p-value<0.05 reject the normality hypothesis

g1=garchFit(~garch(1,1),data=residuals(a.fin2),trace=FALSE,include.me an=TRUE, na.action=na.pass)

summary(g1) g2=garchFit(~garch(1,2),data=residuals(a.fin2),trace=FALSE,include.me an=TRUE, na.action=na.pass)

summary(g2) g3=garchFit(~garch(2,1),data=residuals(a.fin2),trace=FALSE,include.me an=TRUE, na.action=na.pass)

summary(g3) g4=garchFit(~garch(2,2),data=residuals(a.fin2),trace=FALSE,include.me an=TRUE, na.action=na.pass)

summary(g4)

#The best one is Garch(1,1) model which has the smallest AIC.

Unit 4 HW

4.1 Improve the R quantlets on GH (from CRIX directory on quantlet.de) and make excellent graphics that follow Fig 3,4,5,6 of the „Econometrics of CRIX“ paper.

4.2 Make your R code perfect as in the R examples on quantlet.de i.e. make sure that the code is „time independent“ by using actual dimensions of the data that you are collecting from crix.hu-berlin.de RecreateFig 7 from „Econometrics of CRIX“.

4.3 Redo as many figures as you can.

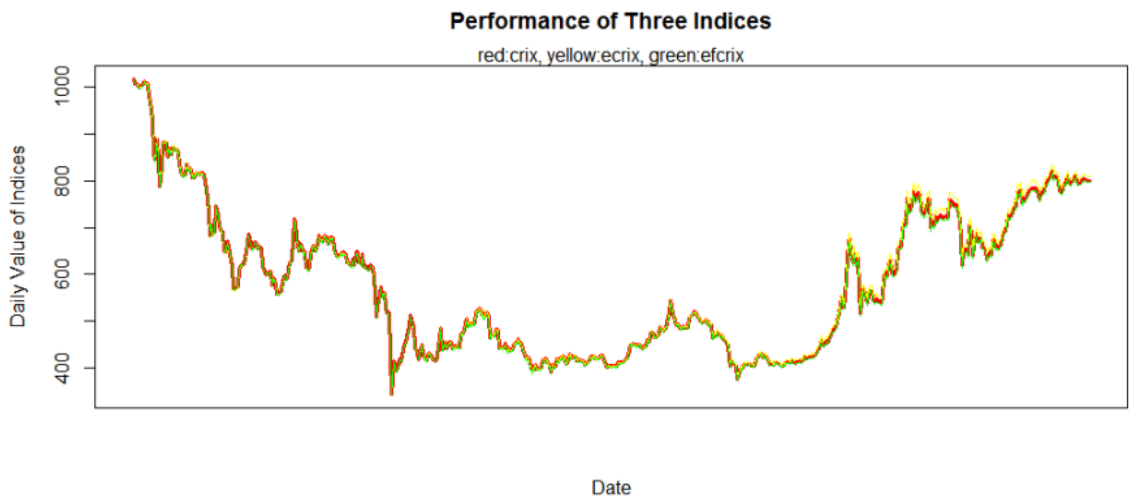
(all this to be done on perfect PPTX slides)

Unit 4 Homework

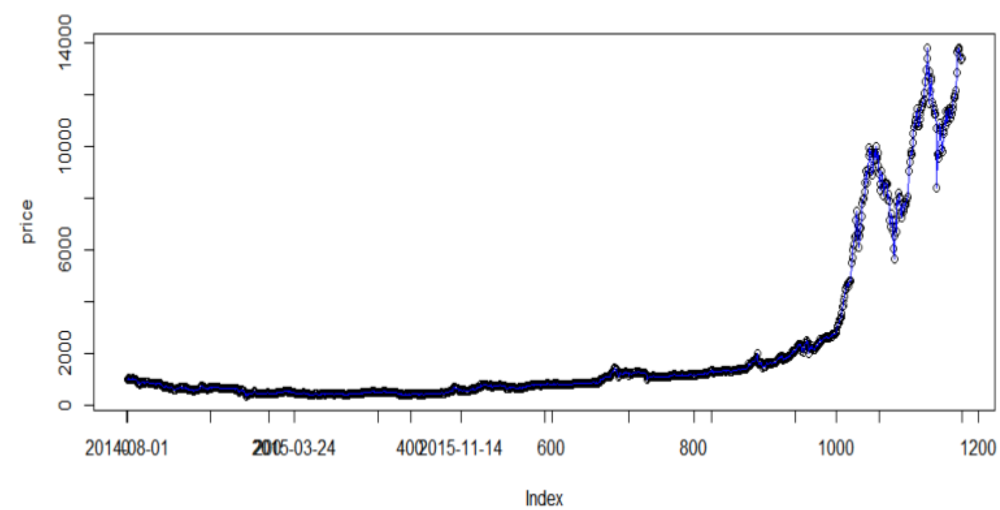
by RUI CHEN 15620161152244

```
#figure 3:crix&ecrix%efcrix
setwd("C:/Users/Administrator/Desktop/workstation")
load("crix.RData")
load("ecrix.RData")
load("efcrix.RData")
plot(crix, type = "l", col = "red", xaxt = "n", lwd = 3, main =
  "Performance of Three Indices",  xlab = "Date", ylab = "Daily Value of
  Indices")
lines(ecrix, col = "yellow")
lines(efcrix, col = "green")
mtext("red:crix, yellow:ecrix, green:efcrix")
```

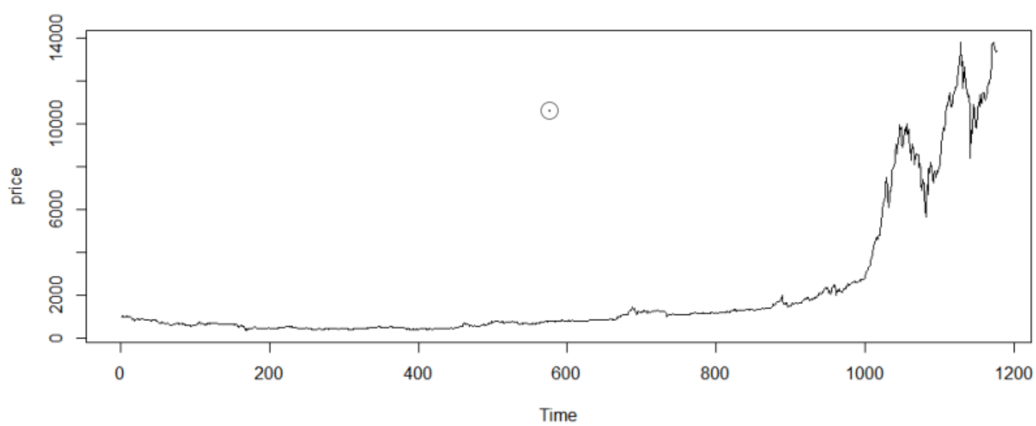
Q1. Figure 3




```
library(rjson)
json_file = "http://crix.hu-berlin.de/data/crix.json" •
json_data = fromJSON(file=json_file)
crix_data_frame=as.data.frame(json_data)
x=crix_data_frame
dim(x)
n=dim(x)  # [1]    1 2354 #
a=seq(1,n[2],2)
b=seq(2,n[2],2)
data=t(x[1,a])
price=t(x[1,b])
plot(price)
lines(price, col = "blue")
```

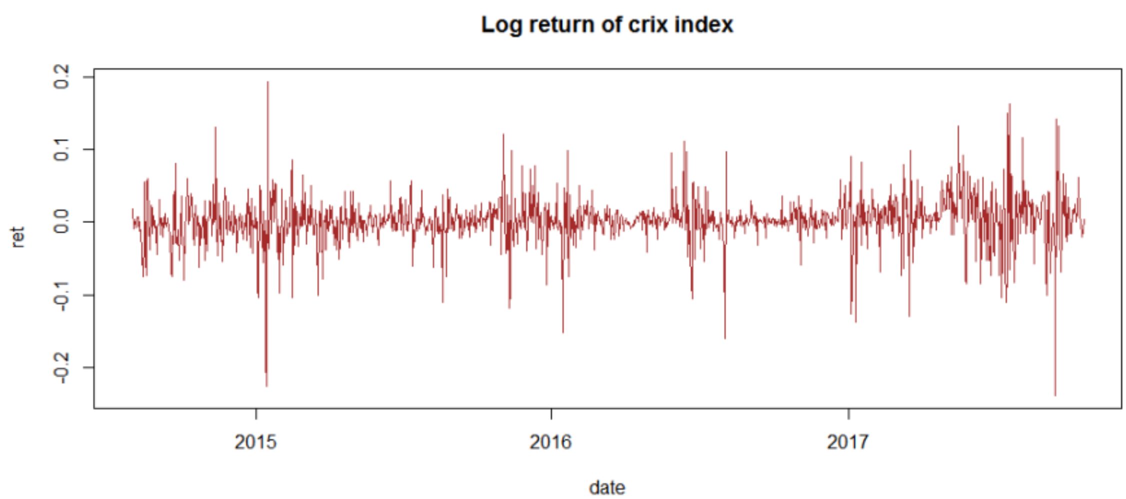


ts.plot(price)



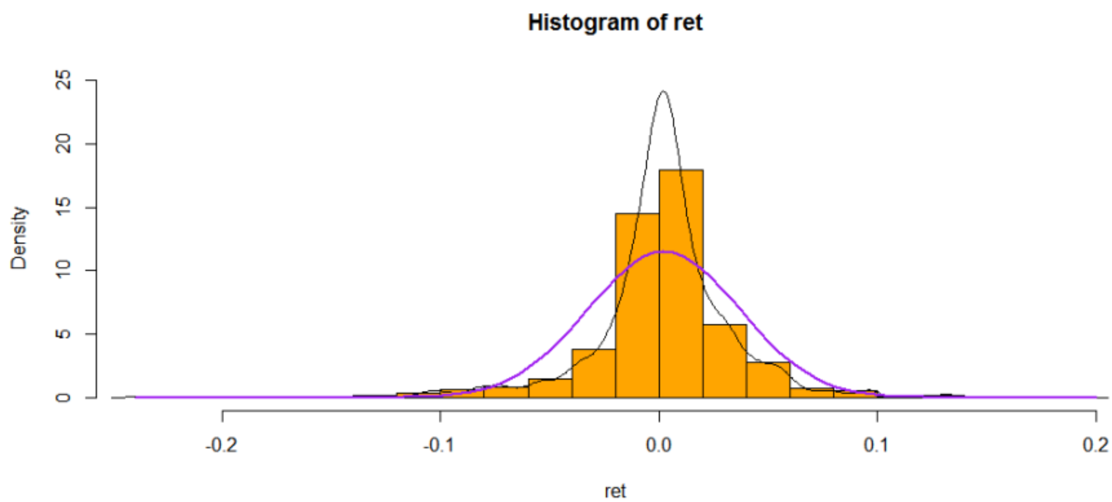
```
#figure4
library(rjson)
json_file = "http://crix.hu-berlin.de/data/crix.json"
json_data = fromJSON(file=json_file)
x = as.data.frame(json_data)
date1=c(json_data[[1]]$date)
for (i in 1:2348){date1[i]=c(json_data[[i]]$date)}
price1=c(json_data[[1]]$price)
for (i in 1:2348){price1[i]=c(json_data[[i]]$price)}
date=date1
price=price1
crix=data.frame(date,price)
date2=date[-1]
ret=diff(log(price))
plot(ret~as.Date(date2),type="l",col="brown",xlab="date",ylab="ret", main="Log
  return of crix index")
```

Q1. Figure 4



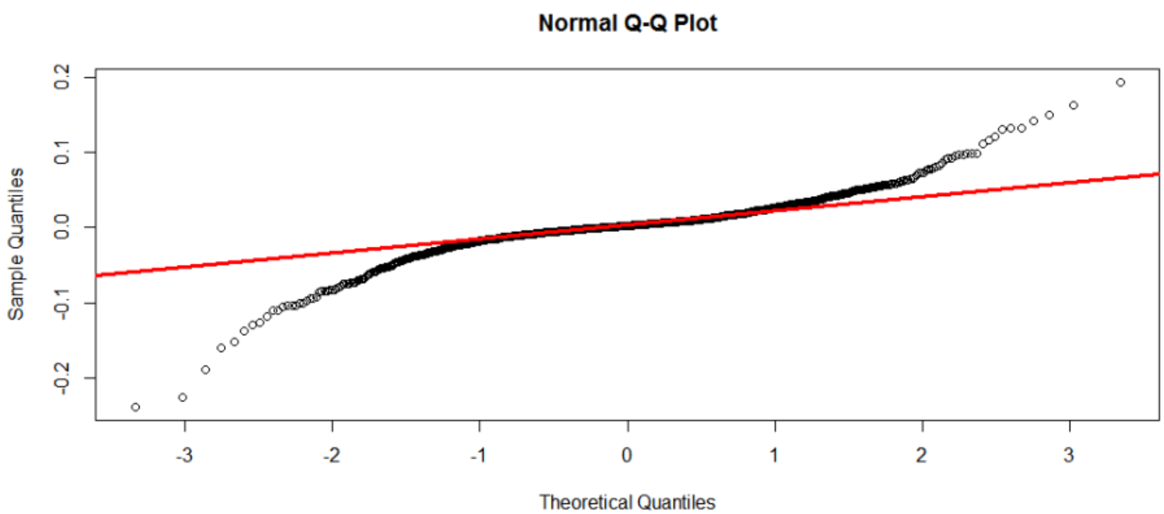
```
#figure5
mean(ret) # [1] 0.002206347 #
var(ret) # [1] 0.001206677 #
sd(ret) # [1] 0.03473726 #
hist(ret, col = "orange", breaks = 20, freq = FALSE, ylim = c(0, 25), xlab = "ret")
lines(density(ret), lwd = 2)
mu = mean(ret)
sigma = sd(ret)
x = seq(-4, 4, length = 100)
curve(dnorm(x, mean = mean(ret), sd = sd(ret)), add = TRUE, col = "purple", lwd
  = 2)
```

Q1. Figure 5



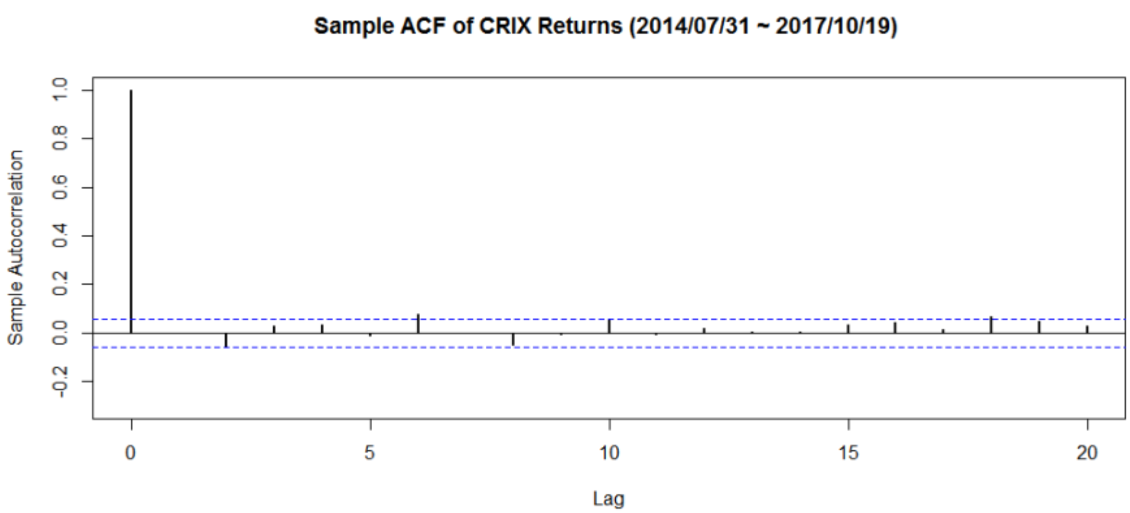
```
qqnorm(ret)
qqline(ret, col = "red", lwd = 3)
```

Q1. Figure 5

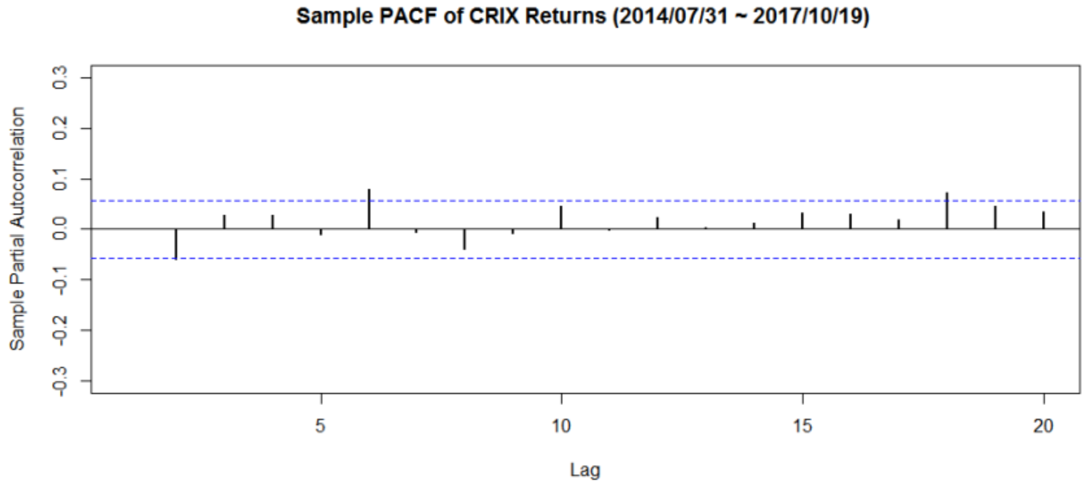


```
#figure6
libraries = c("zoo", "tseries")
autocorr = acf(ret, lag.max = 20, ylab = "Sample Autocorrelation",main
= "Sample ACF of CRIX Returns (2014/07/31 ~ 2017/10/19) ", lwd = 2,
ylim = c(-0.3, 1))
autopcorr = pacf(ret, lag.max = 20, ylab = "Sample Partial
  Autocorrelation", main = "Sample PACF of CRIX Returns
  (2014/07/31 ~ 2017/10/19) ", ylim = c(-0.3, 0.3), lwd = 2)
```

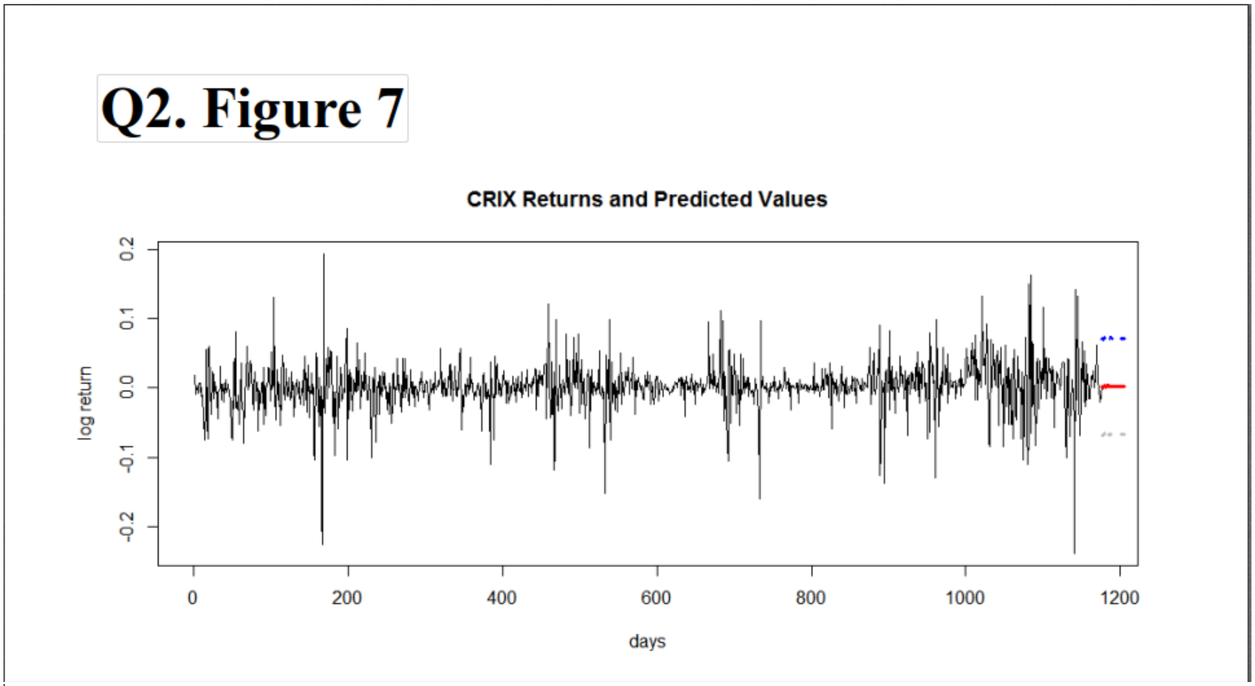
Q1. Figure 6.1



Q1. Figure 6.2




```
#figure7
# arima model
library(caschrono)
library(TTR)
library(forecast)
library(TSA)
par(mfrow = c(1, 1))
auto.arima(ret)
fit202 = arima(ret, order = c(2, 0, 2))
tsdiag(fit202)
fit202 = arima(ret, order = c(2, 0, 2))
crpre = predict(fit202, n.ahead = 30)
dates = seq(as.Date("31/07/2014", format = "%d/%m/%Y"), by = "days", length = length(ret))
plot(ret, type = "l", ylab = "log return", xlab = "days", lwd = 1.5, main = "CRIX Returns and Predicted
    Values")
lines(crpre$pred, col = "red", lwd = 3)
lines(crpre$pred + 2 * crpre$se, col = "blue", lty = 3, lwd = 3)
lines(crpre$pred - 2 * crpre$se, col = "grey", lty = 3, lwd = 3)
```



Thank you!

Unit 5 HW

5.1 Do a word cloud for Shakesspeare’s dramas. Romeo and Julia, Julius Caesar, Hamlet.

5.2 Calculate the histogram of words.

5.3 Map the Shakesspeare words into a dictionary to check its sentiment.

(all this to be done on perfect PPTX slides)

Unit 5 Homework

Rui Chen 15620161152244

Q1

```
rm(list = ls())
#install.packages("RCurl")
#install.packages("XML")
library(RCurl)
library(XML)
url1  = "http://shakespeare.mit.edu/romeo_juliet/full.html"
url2  = "http://shakespeare.mit.edu/julius_caesar/full.html"
url3  = "http://shakespeare.mit.edu/hamlet/full.html"
html1  = readLines(url1, encoding = "UTF-8")
html2  = readLines(url2, encoding = "UTF-8")
html3  = readLines(url3, encoding = "UTF-8")
html1  = htmlParse(html1, encoding = "UTF-8")
html2  = htmlParse(html2, encoding = "UTF-8")
html3  = htmlParse(html3, encoding = "UTF-8")
```

Q1

```
#install.packages("bitops")
#install.packages("stringr")
library(bitops)
library(stringr)
abs1   = lapply(url1, FUN = function(x) htmlParse(x, encoding = "Latin-1"))
abs2   = lapply(url2, FUN = function(x) htmlParse(x, encoding = "Latin-1"))
abs3   = lapply(url3, FUN = function(x) htmlParse(x, encoding = "Latin-1"))
clean_txt = function(x) {
  cleantxt = xpathApply(x, "//body//text()
            [not(ancestor :: script)][ not(ancestor :: style)]
            [not(ancestor :: noscript)] " ,xmlValue)
  cleantxt = paste(cleantxt, collapse="\n")
  cleantxt = str_replace_all(cleantxt, "\n", " ")
  cleantxt = str_replace_all(cleantxt, "\r", "")
  cleantxt = str_replace_all(cleantxt, "\t", "")
  cleantxt = str_replace_all(cleantxt, "<br>", "")
  return(cleantxt)
}
```

Q1

```
cleantxt1 = lapply(abs1,clean_txt)
cleantxt2 = lapply(abs2,clean_txt)
cleantxt3 = lapply(abs3,clean_txt)
vec_abs1 = unlist(cleantxt1)
vec_abs2 = unlist(cleantxt2)
vec_abs3 = unlist(cleantxt3)

###Text Mining
install.packages("tm")
install.packages("SnowballC")
library(tm)
library(SnowballC)
abs1   = Corpus(VectorSource(vec_abs1))
abs2   = Corpus(VectorSource(vec_abs2))
abs3   = Corpus(VectorSource(vec_abs3))
```

```
abs_dtm1  = DocumentTermMatrix(abs1, control = list(
  stemming = TRUE, stopwords = TRUE, minWordLength = 3,
  removeNumbers = TRUE, removePunctuation = TRUE))
abs_dtm2  = DocumentTermMatrix(abs2, control = list(
  stemming = TRUE, stopwords = TRUE, minWordLength = 3,
  removeNumbers = TRUE, removePunctuation = TRUE))
abs_dtm3  = DocumentTermMatrix(abs3, control = list(
  stemming = TRUE, stopwords = TRUE, minWordLength = 3,
  removeNumbers = TRUE, removePunctuation = TRUE))
##WordCloud
instal.packages("ggplot2")
install.packages("wordcloud")
library(ggplot2)
library(wordcloud)
freq1 = colSums(as.matrix(abs_dtm1))
freq2 = colSums(as.matrix(abs_dtm2))
freq3 = colSums(as.matrix(abs_dtm3))
wf1   = data.frame(word=names(freq1), freq=freq1)
wf2   = data.frame(word=names(freq2), freq=freq2)
wf3   = data.frame(word=names(freq3), freq=freq3)
```

Q1

```
#Romeo and Juliet
plot1 = ggplot(subset(wf1, freq>15), aes(word, freq1))
plot1 = plot1 + geom_bar(stat="identity")
plot1 = plot1 + theme(axis.text.x=element_text(angle=45, hjust=1))
plot1
freq1  = colSums(as.matrix(abs_dtm1))
dark2_1 = brewer.pal(6, "Dark2")
wordcloud(names(freq1), freq1, max.words=100, rot.per=0.2, colors=dark2_1)
#Julius Caeser
plot2 = ggplot(subset(wf2, freq>15), aes(word, freq2))
plot2 = plot2 + geom_bar(stat="identity")
plot2 = plot2 + theme(axis.text.x=element_text(angle=45, hjust=1))
plot2
```

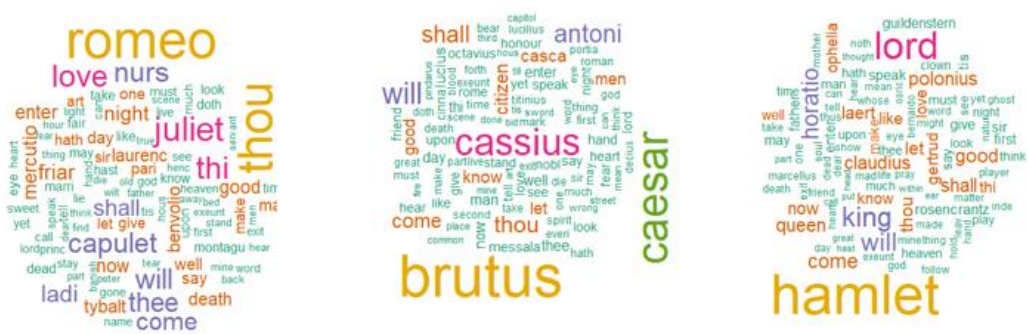
Q1

```
freq2  = colSums(as.matrix(abs_dtm2))
dark2_2 = brewer.pal(6, "Dark2")
wordcloud(names(freq2), freq2, max.words=100, rot.per=0.2, colors=dark2_2)
#Hamlet
plot3 = ggplot(subset(wf3, freq>15), aes(word, freq3))
plot3 = plot3 + geom_bar(stat="identity")
plot3 = plot3 + theme(axis.text.x=element_text(angle=45, hjust=1))
plot3
freq3  = colSums(as.matrix(abs_dtm3))
dark2_3 = brewer.pal(6, "Dark2")
wordcloud(names(freq3), freq3, max.words=100, rot.per=0.2, colors=dark2_3)
```

Q1

```
#Romeo and Juliet
plot1 = ggplot(subset(wf1, freq>15), aes(word, freq1))
plot1 = plot1 + geom_bar(stat="identity")
plot1 = plot1 +
theme(axis.text.x=element_text(angle=45, hjust=1))
plot1
freq1  = colSums(as.matrix(abs_dtm1))
dark2_1 = brewer.pal(6, "Dark2")
wordcloud(names(freq1), freq1, max.words=100,
rot.per=0.2, colors=dark2_1)
#Julius Caeser
plot2 = ggplot(subset(wf2, freq>15), aes(word, freq2))
plot2 = plot2 + geom_bar(stat="identity")
plot2 = plot2 +
theme(axis.text.x=element_text(angle=45, hjust=1))
plot2
freq2  = colSums(as.matrix(abs_dtm2))
dark2_2 = brewer.pal(6, "Dark2")
wordcloud(names(freq2), freq2, max.words=100,
rot.per=0.2, colors=dark2_2)
#Hamlet
plot3 = ggplot(subset(wf3, freq>15), aes(word,
freq3))
plot3 = plot3 + geom_bar(stat="identity")
plot3 = plot3 +
theme(axis.text.x=element_text(angle=45, hjust=1))
plot3
freq3  = colSums(as.matrix(abs_dtm3))
dark2_3 = brewer.pal(6, "Dark2")
wordcloud(names(freq3), freq3, max.words=100,
rot.per=0.2, colors=dark2_3)
```

Q1 figures

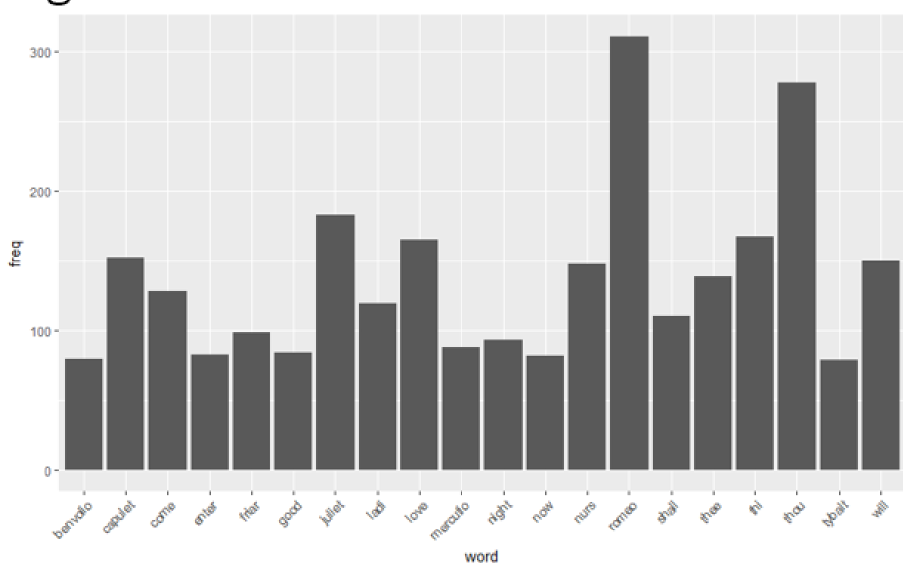


Q2

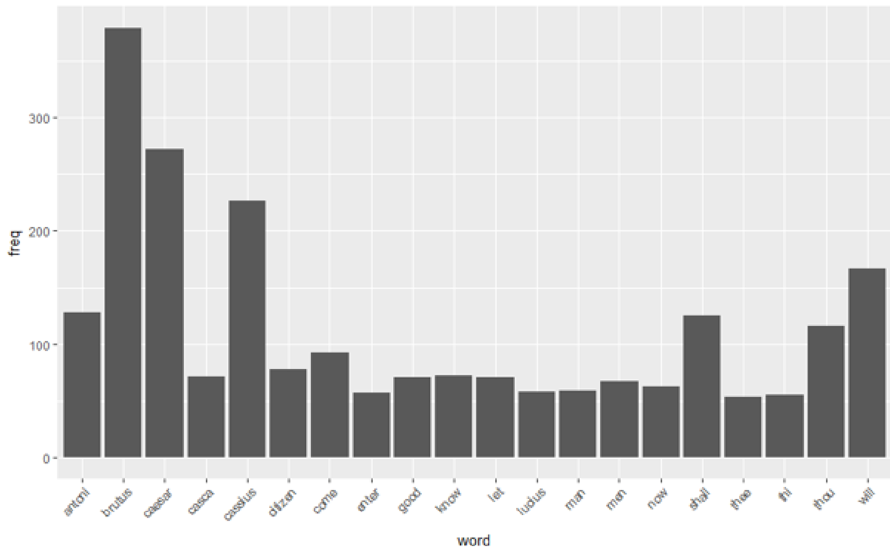
```
#Romeo and Juliet
wf1 <- wf1[order(-wf1$freq),]
wf1 <- wf1[c(1:20),]
p1 = ggplot(subset(wf1, freq > 15), aes(word, freq))
p1 = p1 + geom_bar(stat = "identity")
p1 = p1 + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p1
#Julius Caeser
wf2 <- wf2[order(-wf2$freq),]
wf2 <- wf2[c(1:20),]
p2 = ggplot(subset(wf2, freq > 15), aes(word, freq))
p2 = p2 + geom_bar(stat = "identity")
p2 = p2 + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p2
```

```
#Hamlet
wf3 <- wf3[order(-wf3$freq),]
wf3 <- wf3[c(1:20),]
p3 = ggplot(subset(wf3, freq > 15), aes(word, freq))
p3 = p3 + geom_bar(stat = "identity")
p3 = p3 + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p3
```

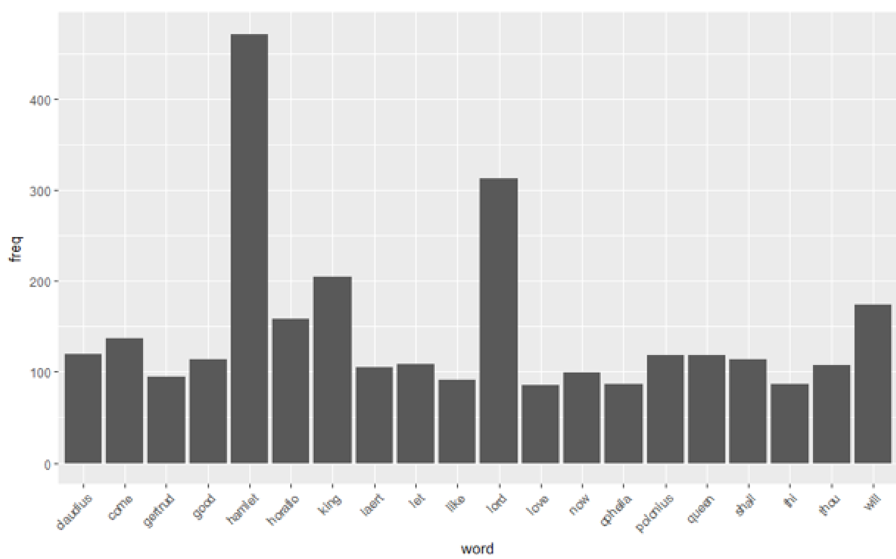
Q2 figures

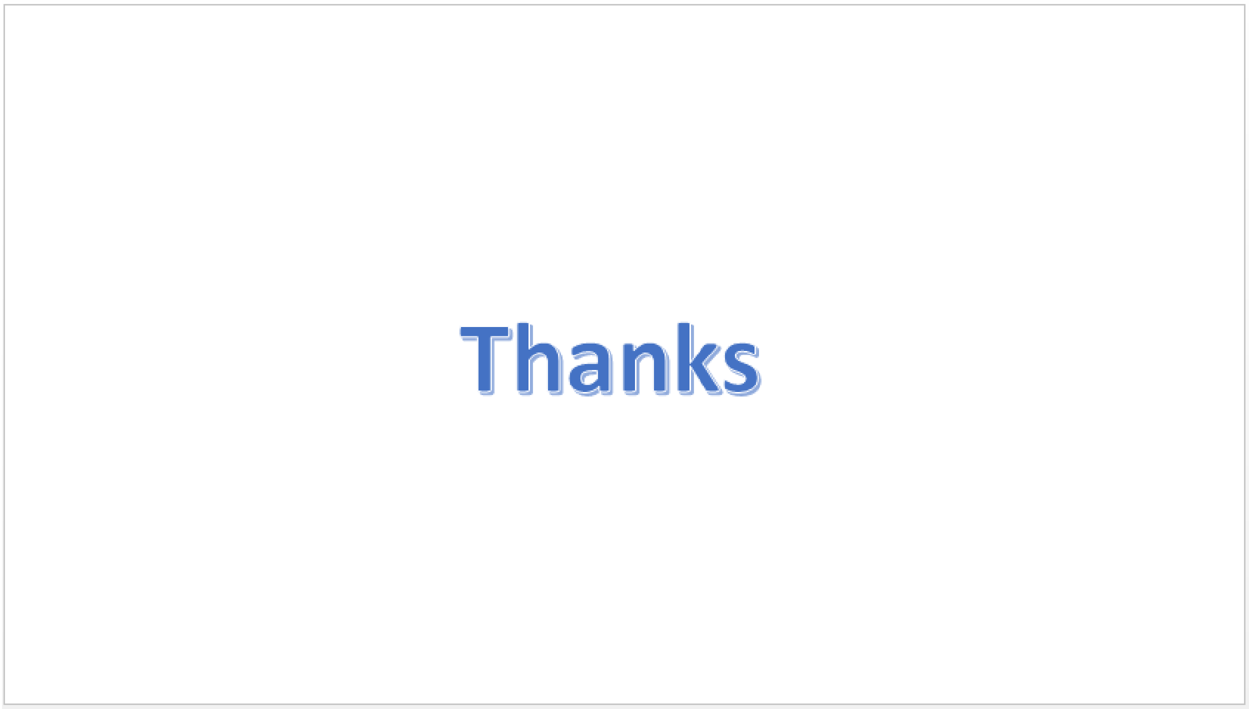


Q2 figures



Q2 figures





Final EXAM

Collect all HW s in one word file and leave it on GH for evaluation.