Big data homework

**Hanjun Hua 27720161153020**

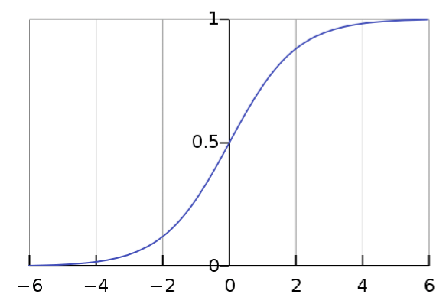
**HW1-1 Memory of personal computer**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| date | number | date | number | date | number |
| 1971 | 1 | 1990 | 1 | 2000 | 22 |
| 1972 | 1 | 1991 | 1 | 2001 | 27 |
| 1974 | 1 | 1992 | 4 | 2002 | 44 |
| 1976 | 1 | 1993 | 1 | 2003 | 31 |
| 1978 | 1 | 1994 | 3 | 2004 | 27 |
| 1979 | 1 | 1995 | 3 | 2005 | 16 |
| 1982 | 2 | 1996 | 2 | 2006 | 29 |
| 1985 | 1 | 1997 | 5 | 2007 | 44 |
| 1988 | 1 | 1998 | 12 | 2008 | 59 |
| 1989 | 1 | 1999 | 21 |  |  |

Table1 The number of annual production for the memory of personal computers

**HW1-2 Logistic Regression**

1. Logistic regression model is a classification model. There are many similarities between the logistic regression model and multiple linear regression, the biggest difference is the dependent variable. Both of these are generalized linear models.
2. Definition:Logistic regression model is the following conditional probability distribution:
3. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probability regression using similar techniques, with the latter using a cumulative normal distribution curve instead.
4. Logistic regression is mainly applied in epidemiology. Such as to explore the risk factors of certain diseases, and predict the probability of a disease according to the risk factors. In addition, it can be used survival analysis, hydrology and economics.



**HW 2**

**# HW 2-1 & 2-2 #**

year<-c(1971,1972,1974,1976,1978,1979,1982,1985,1988,1989,1990,1991,1992,1993, 1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008)

number<-c(1,1,1,1,1,1,2,1,1,1,1,1,4,1,3,3,2,5,12,21,22,27,44,31,27,16,29,44,59)

plot(year,number,type = "b",

col="black",main = "The history of computer memory", sub = "This is the change in the number of types of computer memory", xlab = "year",ylab = "The number of memory")

barplot(number, xlab = "year",ylab ="The number of memory ")

**# HW 2-3 #**

lambda=4

x=6

dpois(x,lambda)

lambda=5

x=0

dpois(x,lambda)

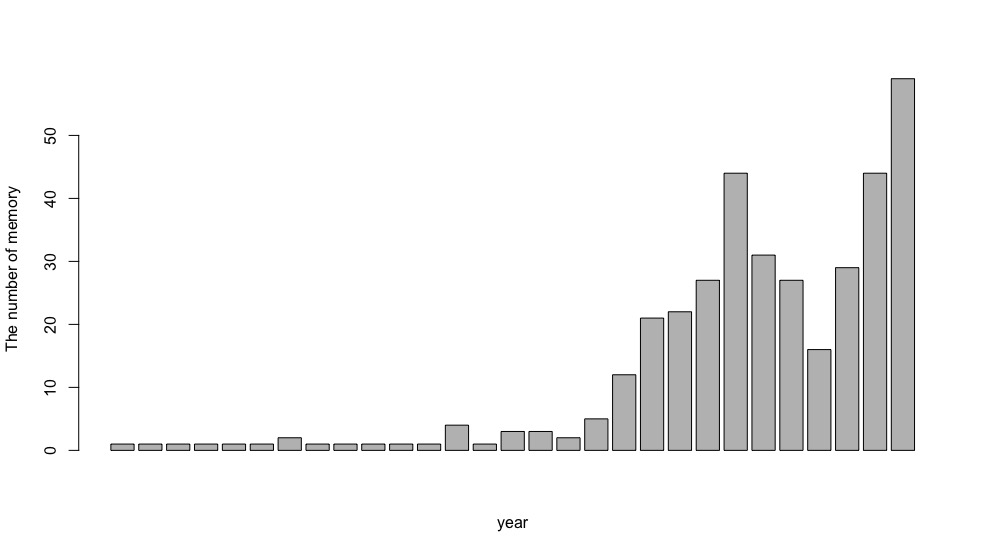


Fig1 The result of hw2-1 and hw2-2

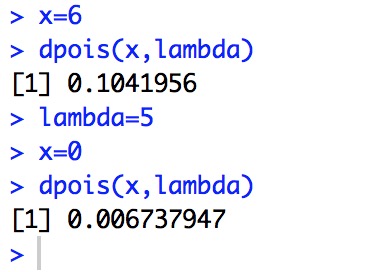


Fig2 The result of hw2-3

**HW3-1**

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

library("digest")

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

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

**HW3-2 Digital Signature Algorithms**

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.

Key generation included parameter generation and per-user keys.

Key generation has two phases. The first phase is a choice of algorithm parameters which may be shared between different users of the system, while the second phase computes public and private keys for a single user.

• Let H be the hashing function and m the message:

• Generate a random per-message value k where 1<k<q

• Calculate r=(gk mod p) mod q

• In the unlikely case that r=0, start again with a different random k • Calculate s=k−1(H(m)+xr) mod q

• In the unlikely case that s=0, start again with a different random k • The signature is (r,s)

• Reject the signature if 0<r<q or 0<s<q is not satisfied • Calculate w=s−1 mod q

• Calculate u1=H(m)⋅w mod q

• Calculate u2=r⋅w mod q

• Calculate v=(gu1yu2 mod p) mod q

• The signature is invalid unless v=r

DSA is similar to the ElGamal signature scheme.

With DSA, the entropy, secrecy, and uniqueness of the random signature value k are critical. It is so critical that violating any one of those three requirements can reveal the entire private key to an attacker. Using the same value twice (even while keeping k secret), using a predictable value, or leaking even a few bits of k in each of several signatures, is enough to reveal the private key x.

**HW3-3**

library("RJSONIO")

number<-c(1010,1011,1012,1013,1014,1015,1016)

name<-c("Alice","Bob","Cindy","David","Ella","Frank","Gres")

da<-as.matrix(data.frame(number,name))

s<-cat(toJSON(da))

writeLines(da,"name.json")

**HW3-4**

#download data#

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)

crix\_data\_frame=as.data.frame(json\_data)

crix\_data\_frame\_t<-t(crix\_data\_frame)

time<-crix\_data\_frame\_t[seq(1,2350,by=2)]

price<-crix\_data\_frame\_t[seq(2,2350,by=2)]

crix\_data\_frame<-cbind(time,price)

time\_series<-ts(data=price,start =c(2014,7,31),frequency = 365)

plot(time\_series)

#ARMA or ARIMA?#

library(tseries)

adf.test(time\_series)

#Since p-value greater than printed p-value,we can't reject the hypothesis#

#the time series(time\_series) is not stationary#)

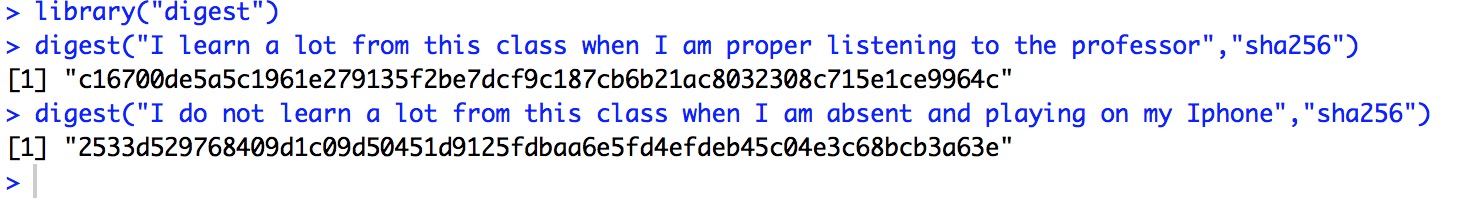


Fig3 The result of hw3-1

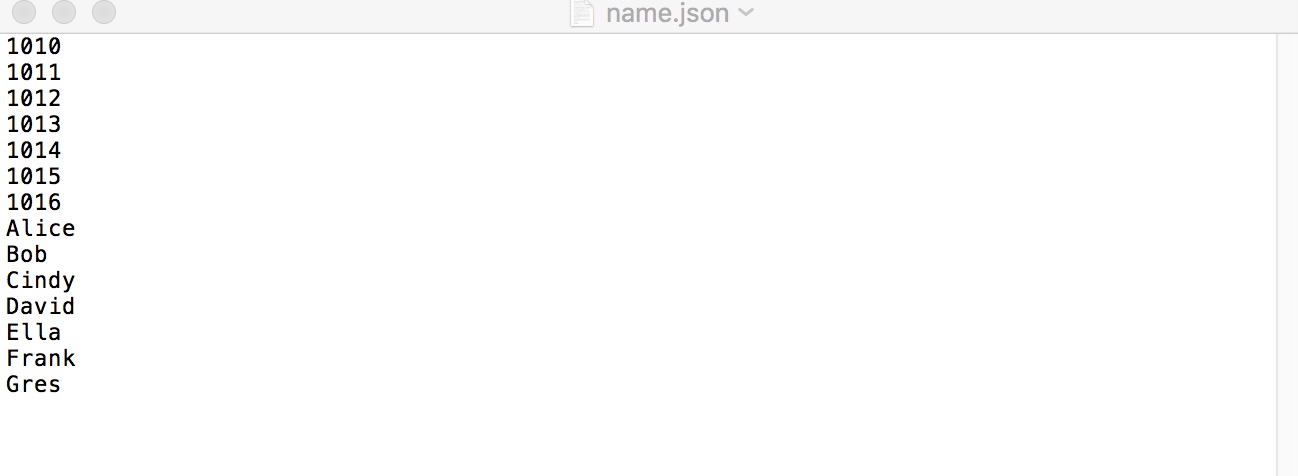


Fig4 The result of hw3-3

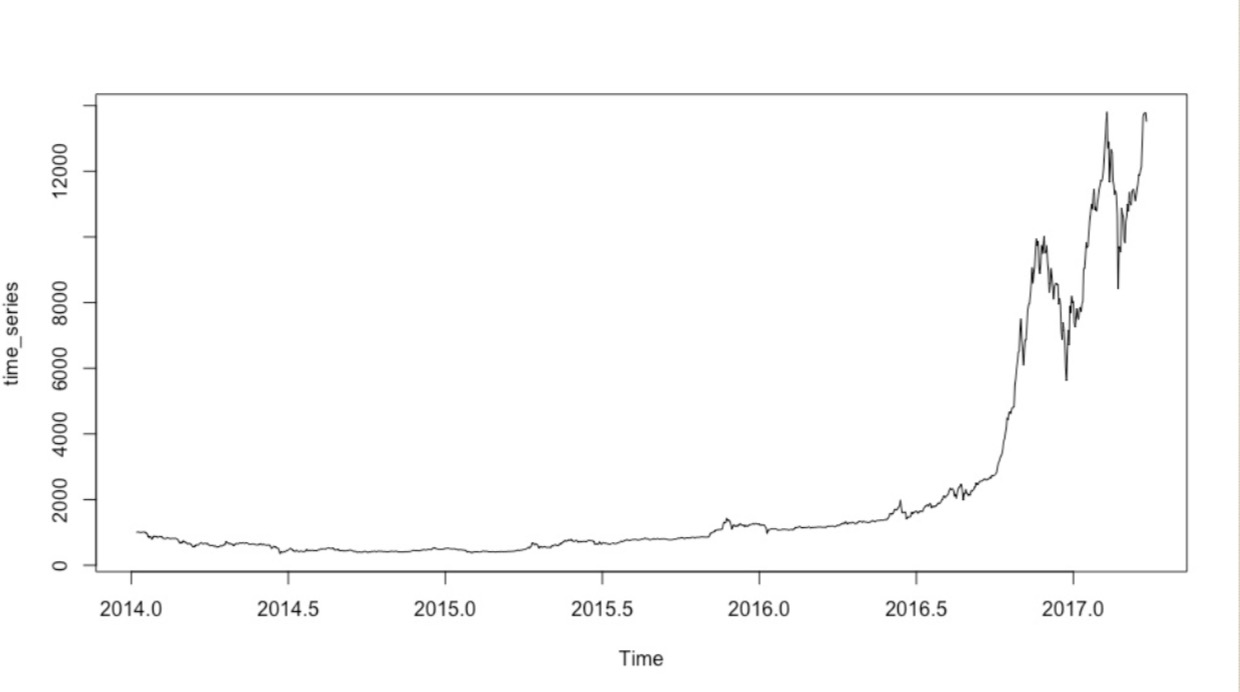


Fig5 The result of hw3-4

**HW4**

**HW4-1**

#download data#

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)

crix\_data\_frame=as.data.frame(json\_data)

#polt of time series #

par(mfrow=c(3,2))

crix\_data\_frame\_t<-t(crix\_data\_frame)

time<-crix\_data\_frame\_t[seq(1,2350,by=2)]

price<-as.numeric(crix\_data\_frame\_t[seq(2,2350,by=2)])

crix\_data\_frame<-cbind(time,price)

time\_series<-ts(data=price,start =c(2014,7,31),frequency = 365)

plot(time\_series,main="The time series of price")

#plot of return #

ret<-diff(log(price))

plot(ts(ret,start= c(2014,7,31),frequency = 365),main="The time series of ret")

#histogram of price and ret#

hist(price,breaks=50,col = "red",freq = FALSE,xlab="price",main = "Histogram of price")

lines(density(price),col="black",lwd=2)

hist(ret,breaks = 40,col = "blue",freq = FALSE,main = "Histogram of price")

lines(density(ret),col="black",lwd=2)

x=seq(-4,4,length=100)

curve(dnorm(x,mean=mean(ret),sd=sd(ret)),add=TRUE,col="red",lwd=2)

#normal qq plot#

qqnorm(ret)

qqline(ret,col="blue",lwd=3)

par(mfrow=c(1,1))

qqnorm(fit$residuals)

qqline(fit$residuals)

Box.test(fit$residuals,type = "Ljung-Box")

#model forecast#

forecast(fit,5)

plot(forecast(fit,5),xlab = "time",ylab = "price")

**HW4-2&4-3**

# HW4-2 and HW4\_3 #

library(forecast)

library(tseries)

#ACF and PACF#

par(mfrow=c(2,2))

Acf(price)

Pacf(price)

Acf(d\_price)

Pacf(d\_price)

#fit model#

fit<-arima(time\_series,order = c(0,2,3))

fit

accuracy(fit)

#model test#

par(mfrow=c(1,1))

qqnorm(fit$residuals)

qqline(fit$residuals)

Box.test(fit$residuals,type = "Ljung-Box")

#model forecast#

forecast(fit,5)

plot(forecast(fit,5),xlab = "time",ylab = "price")

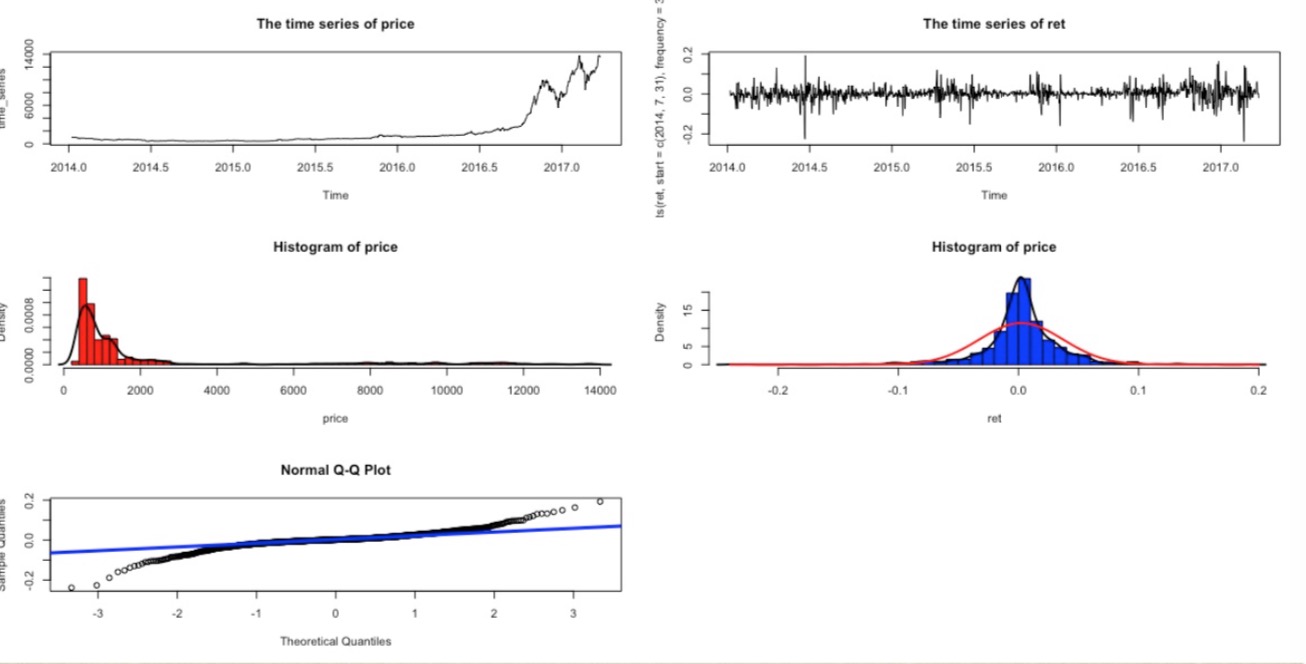


Fig6 The result of hw4-1

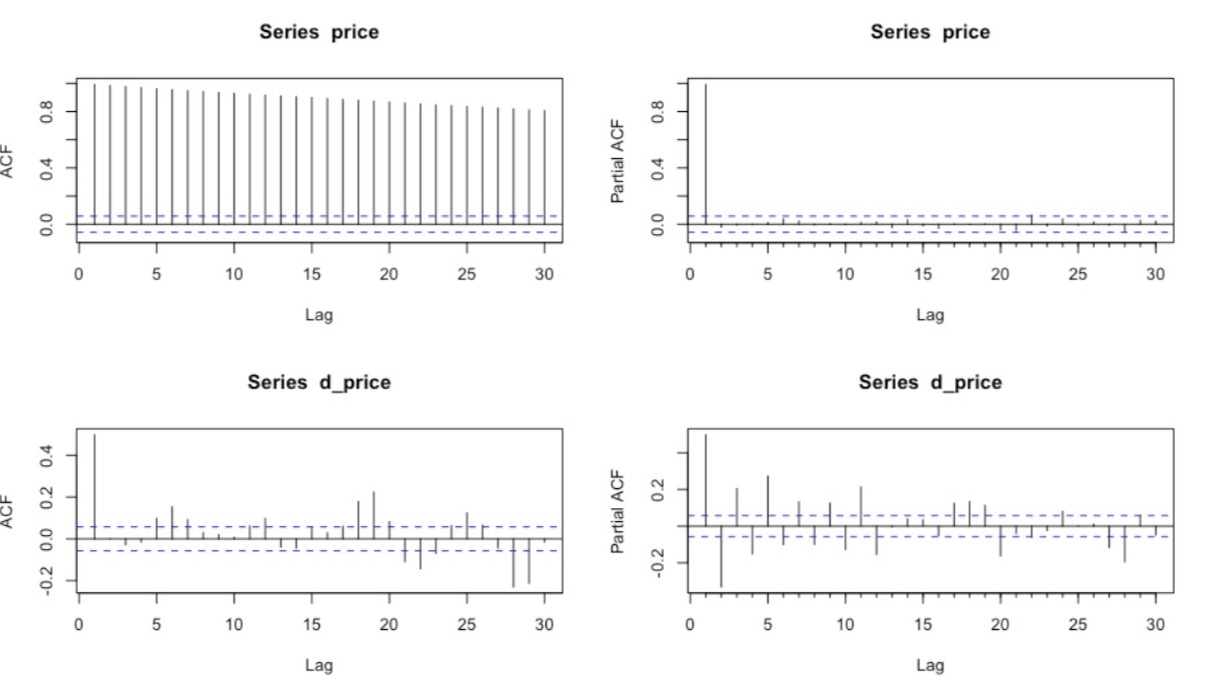


Fig6 The result of hw4-2&4-3

**HW5**

vec\_abs<-c("ROMEO: He jests at scars that never felt a wound. But,soft! what light through yonder window breaks? It is the east，and Juliet is the sun．Arise, fair sun, and kill the envious moon, Who is already sick and pale with grief, That thou her maid art far more fair than she: Be not her maid, since she is envious；Her vestal livery is but sick and green And none but fools do wear it; castit off. It is my lady, O, it is my love! O, that she knew she were! She speaks yet she says nothing: what of that? Her eye discourses, I will answer it．I am too bold, 'tis not to me she speaks: Two of the fairest stars in all the heaven, Having some business, do entreat her eyes To twinklein their spheres till they return. What if her eyes were there, they in her head? The brightness of her cheek would shame those stars, As daylight doth a lamp; her eyes in heavenWould through the airy region stream so bright That birds would sing and think it were not night. See, how she leans her cheek upon her hand! O, that I were a glove upon that hand, That I might touch that cheek!" ,

"JULIET: Ay me! ",

"ROMEO: She speaks: O, speak again, bright angel!for thou art As glorious to this night, being o'er my head As is a winged messenger of heaven Unto the white-upturned wondering eyes Of mortals that fall back to gazeon himWhen he bestrides the lazy-pacing clouds And sails upon the bosomof the air. ",

"JULIET: O Romeo, Romeo! whereforeart thou Romeo? Deny thy father and refuse thy name; Or, if thou wilt not，be but sworn my love, And I'll no longer be a Capulet.",

"ROMEO: Shall I hear more，or shall I speak at this?",

"JULIET: 'Tisbut thy name that is my enemy; Thou art thyself，though not a Montague. What's Montague? it is nor hand, nor foot, Nor arm, nor face, nor any other part Belonging to a man. O, be some other name! What's in a name? that which we call a roseBy any other name would smell as sweet; So Romeo would, were he not Romeo call'd, Retain that dear perfection which he owes Without that title. Romeo, doffthy name, And for that name which is no part of thee Take all myself.",

"ROMEO: I take thee at thy word: Call me but love, and I'll be new baptized, HenceforthI never will be Romeo. ",

"JULIET: What man art thou that thus bescreen'din night So stumblest on my counsel? ",

"ROMEO: By a nameI know not haw to tell thee who I am：My name, dear saint, is hateful to myself, Because it is an enemy to thee, Had I it written,I would tear the word. ",

"JULIET: My ears have not yet drunk a hundred words Of that tongue's utterance, yet I know the sound:Art thou not Romeo and a Montague? ",

"ROMEO: Neither，fair saint, if either thee dislike. ",

"JULIET: How camest thou hither，tell me，and wherefore? The orchard walls are high and hard to climb, And the place death, considering who thou art, If any of my kinsmen find thee here."

)

library(NLP)

library(tm)

library(SnowballC)

abs = Corpus(VectorSource(vec\_abs))

abs\_dtm = DocumentTermMatrix(abs, control = list(

stemming = TRUE, stopwords = TRUE, minWordLength = 3,

removeNumbers = TRUE, removePunctuation = TRUE))

dim(abs\_dtm)

inspect(abs\_dtm)

findFreqTerms(abs\_dtm, 3)

removeSparseTerms(abs\_dtm, 0.5)

inspect(removeSparseTerms(abs\_dtm, 0.5))

library(ggplot2)

library(RColorBrewer)

library(wordcloud)

freq = colSums(as.matrix(abs\_dtm))

wf = data.frame(word=names(freq), freq=freq)

#HW5.2

plot = ggplot(subset(wf, freq>1), aes(word, freq))

plot = plot + geom\_bar(stat="identity")

plot = plot + theme(axis.text.x=element\_text(angle=45, hjust=1))

plot

#HW5.1

dark2 = brewer.pal(6, "Dark2")

wordcloud(names(freq), freq, max.words=100, rot.per=0.2, colors=dark2)