STAT542 HW1

Xiruo Li (xiruoli2)

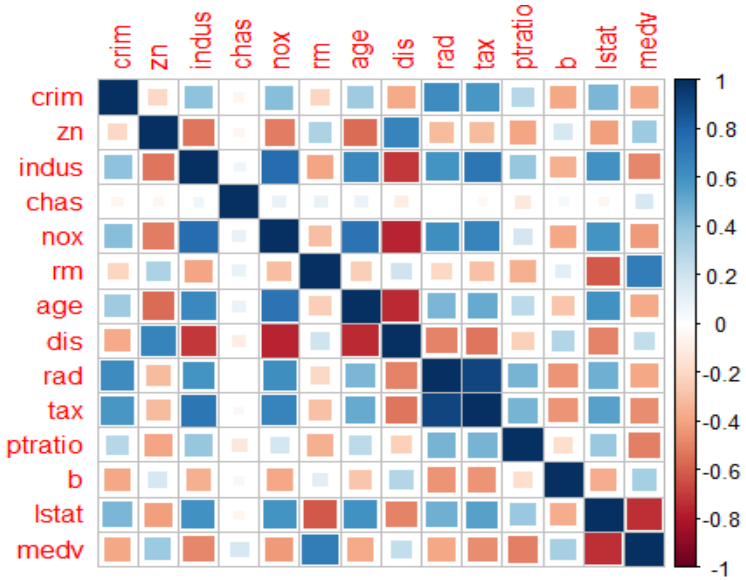
Question 1

(a)

Mean and variance of data:

```
##                       crim          zn        indus chas           nox
## mean            3.6135236   11.363636   11.1367787   NA 5.546951e-01
## var            73.9865782  543.936814   47.0644425   NA 1.342764e-02
##                        rm          age          dis          rad
## mean         6.284634e+00 6.857490e+01 3.795043e+00    9.5494071
## var          4.936709e-01 7.923584e+02 4.434015e+00   75.8163660
##                       tax      ptratio            b        lstat
## mean         4.082372e+02 1.845553e+01 3.566740e+02   12.6530632
## var          2.840476e+04 4.686989e+00 8.334752e+03   50.9947595
##                      medv
## mean         2.253281e+01
## var          8.458672e+01
```

Correlation matrix:




```
#From the stat.desc, we can see the basic statistics of variables, like
 mean and variance I show above.
#Also, from the correlation table, we can find that rad and tax has a h
igh correlation (0.91), which may cause the multicollinearity. So, we m
ay drop either one.
```

(b)

```
#Best subset using BIC: medv=-0.522lstat+3.801rm-0.946ptratio-1.492dis-
17.376nox+2.718chas+0.009b+0.045zn-0.108crim+0.299rad-0.011tax, which h
as the minimum BIC:-608.0353
```

(c)

```
# forward method using AIC, medv=-0.522lstat+3.801rm-0.946ptratio-1.492
dis-17.376nox+2.718chas+0.009b+0.045zn-0.108crim+0.299rad-0.011tax,whic
h has the minimum AIC: 1585.76
# backward method using Cp, medv=-0.522lstat+3.801rm-0.946ptratio-1.492
dis-17.376nox+2.718chas+0.009b+0.045zn-0.108crim+0.299rad-0.011tax,whic
h has the minimum Cp:10.11455
# Compared these three model, they have same variables and parameters.
```

(d)

```
# Backward and forward yield a single model and they are faster than be
st subset method since they add or remove one predictor at a time and d
o not access all models. But, they may miss the best model.
# For the best subset model, since it access all of possible models, it
 will give us the best model for each number of parameters. But,if the
number of parameter is too big, it may take too much time.
# If I got different results among these method, I will choose the mode
l from best subset model, because it go over all of possible cases.
```

(e)

```
# All of these three criteria make a trade off between goodness of fit
and complexity. BIC makes more penalty on the bigger model and it pick
a smaller model than AIC. Also, Cp performs similarly to AIC.
# If I got different results among these criteria, I'd like to choose B
IC since I prefer simpler model.
```

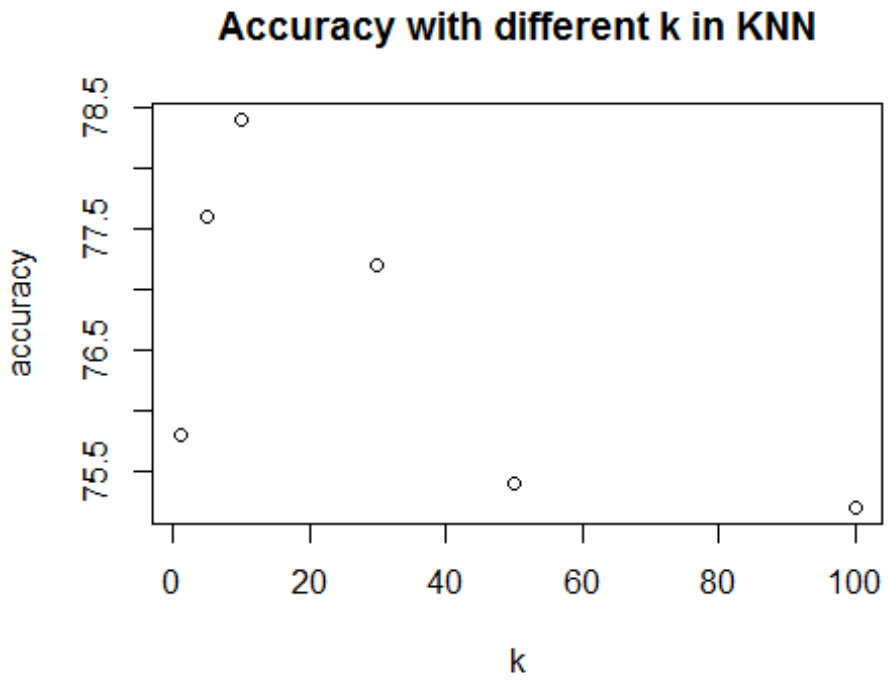
Question2

(a)

```
# Summary: There are 60000 and 10000 observations in the training and t
est data. Each observation has 784 pixels and 1 label.
# Research goal: use KNN algorithm to fit the train data and predict th
e label of test data according to the pixel value.
```

(b)&(c)&(d)

```
## [1] "The accuracy is: 75.8 when k= 1"   "The accuracy is: 77.6 when
k= 5"
## [3] "The accuracy is: 78.4 when k= 10"  "The accuracy is: 77.2 when
k= 30"
## [5] "The accuracy is: 75.4 when k= 50"  "The accuracy is: 75.2 when
k= 100"
```




```
##part(b)
#I choose Euclidean distance as the distance in KNN. Besides, I have fu
nctions for KNN algorithm and Accuracy.
#To get the greatest frequency among k labels, I use the function "sort
(table())".When there is a tie, this function rank the smallest label a
s the 1st.
#But, I found this program run too slowly. Thus, I make some modificati
ons, which will be illustrated in part (d).

##part(d)
#(1)I randomly pick 6000 samples from train set and 500 samples from te
st set, which are my new train set and test set. Because the original d
ataset is too large to run in a short time. And the change of sample si
ze to such a degree will not decrease much accuracy.
#(2)I use PCA and pick up the first 10 components. This method will red
uce the dimension in a large degree.

##part(c)
#I use k=1,5,10,30,50,100. And I found that the accuracy increase first
ly and decrease later. When k increase, the model complexity decrease(b
ias increase and variance decrease). There is a bias-variance trade off
 to get the best performance.
#When k=10, I got the optimal performance(accuracy=78.4%). In this case
, the d.f=n/k=6000/10=600.
```