Midterm Exam

36-350, Data Mining

In class, 15 October 2008

When doing calculations, show your work. No calculators are needed, or allowed. Ask me if you need extra paper. There are three numbered questions, all equally weighted. You are not expected to do all parts of all problems.

Table of natural logarithms (to three significant digits). Remember that $\log 1/n = -\log n$, and $\log ab = \log a + \log b$.

n	$\log n$
1	0
2	0.693
3	1.10
4	1.39

`R` reminders:

```
lm(y~x1+x2) # Linearly regress y on x1 and x2, but not their interactions
sapply(x,f) # Apply the function f to every element in the vector x
apply(x,1,f) # Apply the function f to every row in the array x
apply(x,2,f) # Apply f to the columns of x
kmeans(x,k) # Use k-means to divide the data matrix x into k clusters
prcomp(x,scale.=TRUE) # Principal component analysis of data matrix,
                      # with columns scaled to variance 1
cor(x) # Correlation matrix of a data frame x
cov(x) # Covariance matrix
```

1. *Wine recommendation* An organization of wine ~~snobs~~ enthusiasts gathers data on the wines its members have in their collections. They use the **box-of-wines** representation, so the data consists of a list of wine types for each member, like so:

```
Aligheri,D: "L'Inferno 00", "Beata Beatrix 83", "Averno 01", "Ad Astra 02"
Maro,PV: "L'Inferno 00", "Averno 01", "Ad Astra 02", "Maecenas Estates n.d."
Montresor,E: "Valdemar Revenant 03", "L'Inferno 00"
Fortunato,A: "Pym Amontillado 02", "Ligeia 99"
Rumpole,J: "Chateau Plonk 97", "Chateau Plonk 98", "Chateau Plonk 99",
           "Chateau Plonk 00", "Chateau Plonk 01", "Chateau Plonk 02"
```

   There will, however, be many more than five members.

   We want to use this data to build a system to recommend new wines to people, based on the ones they already like.

   (a) (8 points) Explain how to convert this data into a data frame with one row for each member, and a binary value for each feature. Be clear about what the features are, how they relate to wines, and how you would determine the number of features from the data.

   (b) (10 points) Explain how to find the k wine-drinkers most similar to a given member. Be explicit about how you would calculate similarity.

   (c) (5 points) Explain how to calculate "inverse drinker frequency". What is the IDF of the wine "L'Inferno 00" among the 5 drinkers above? Do you think IDF would be useful in finding similar drinkers?

   (d) (10 points) Describe a procedure to recommend up to m wines to a customer based on the wines enjoyed by the k most-similar people in the data base. Explain how your procedure ensures that these wines are new to the customer, and how it selects m recommendations if it comes up with more than m candidates.

   (e) (Extra credit; 10 points) Explain how to modify your procedure to use latent semantic indexing.

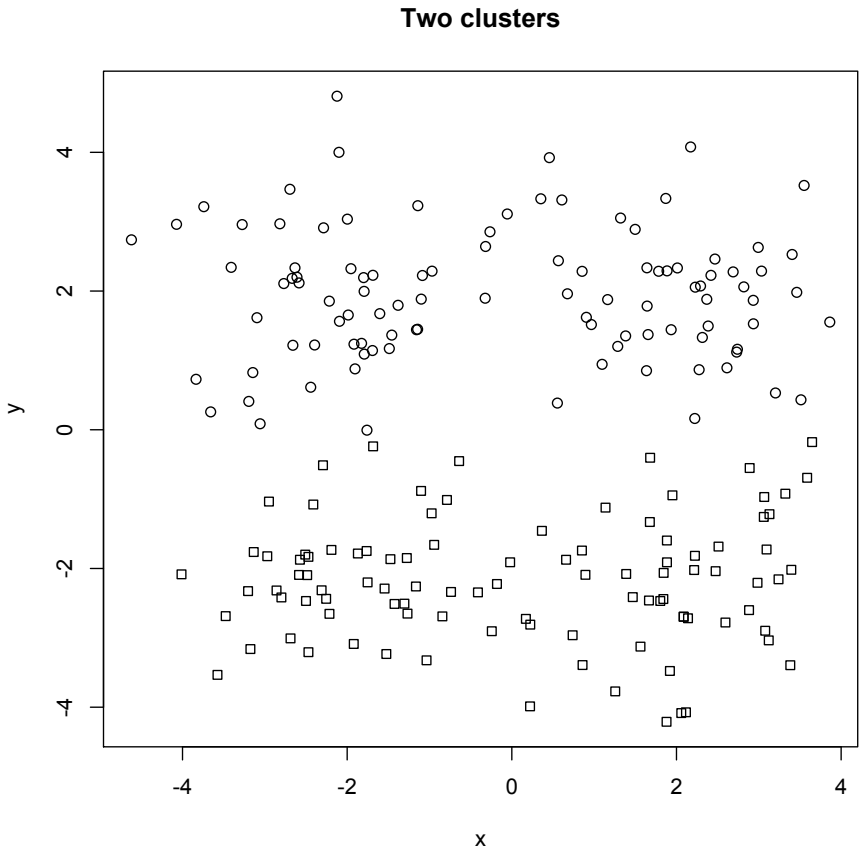
2. *k-Means Clustering*

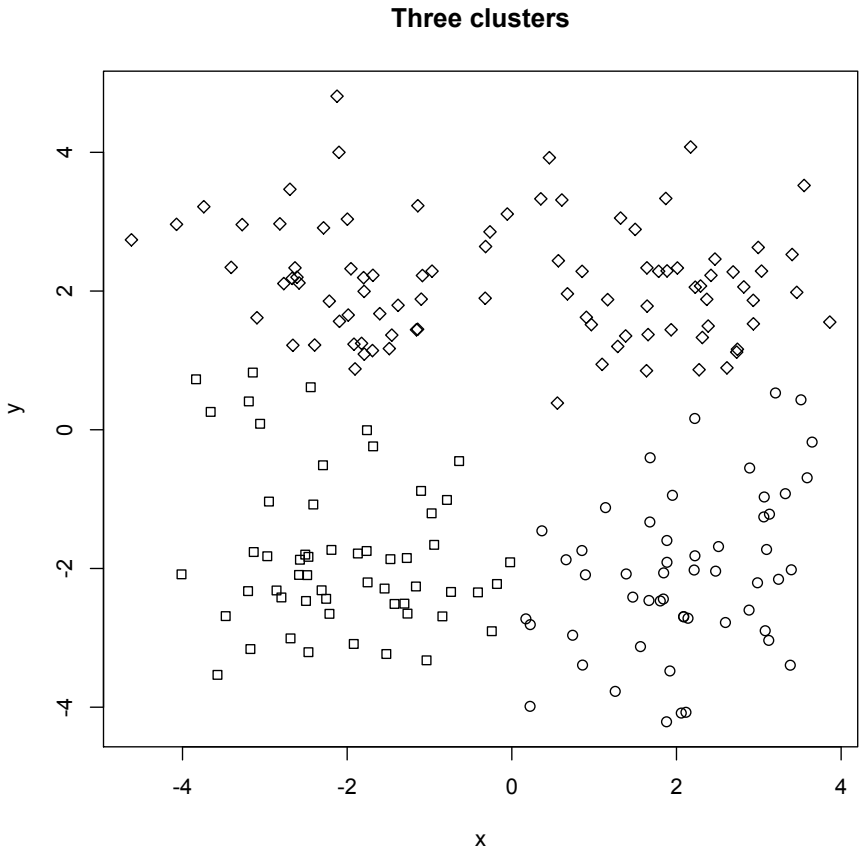
   (a) (15 points) Explain what the k-means clustering algorithm is. You do not need to write code, but give a precise verbal description which someone could turn into code.

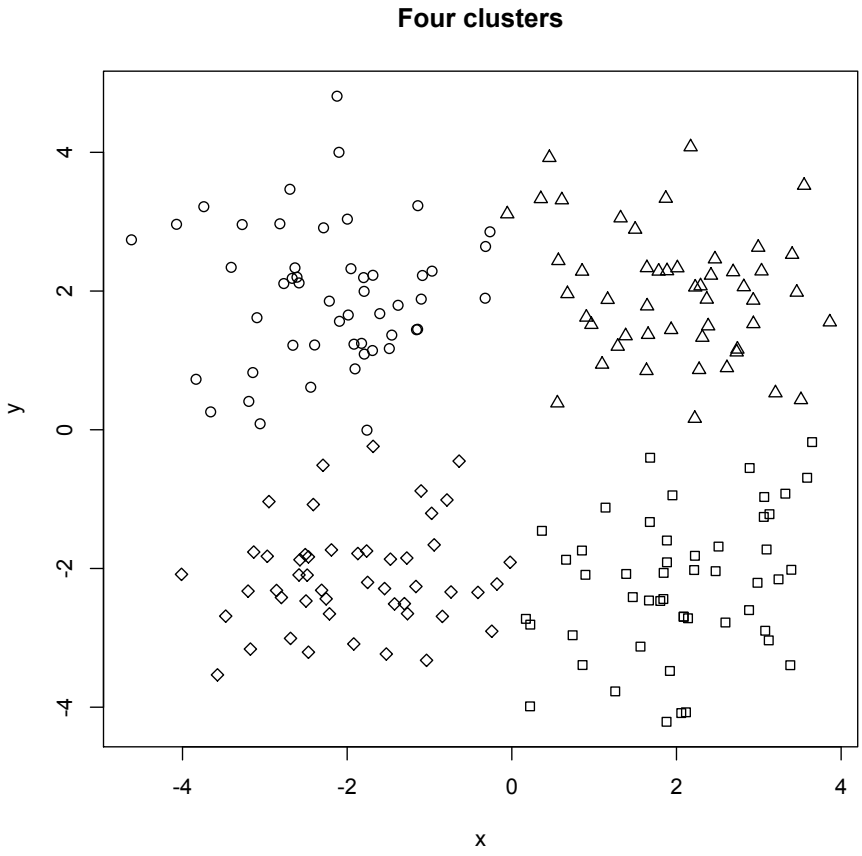
   (b) (5 points) Can k-means ever give results which contain more or less than k clusters?

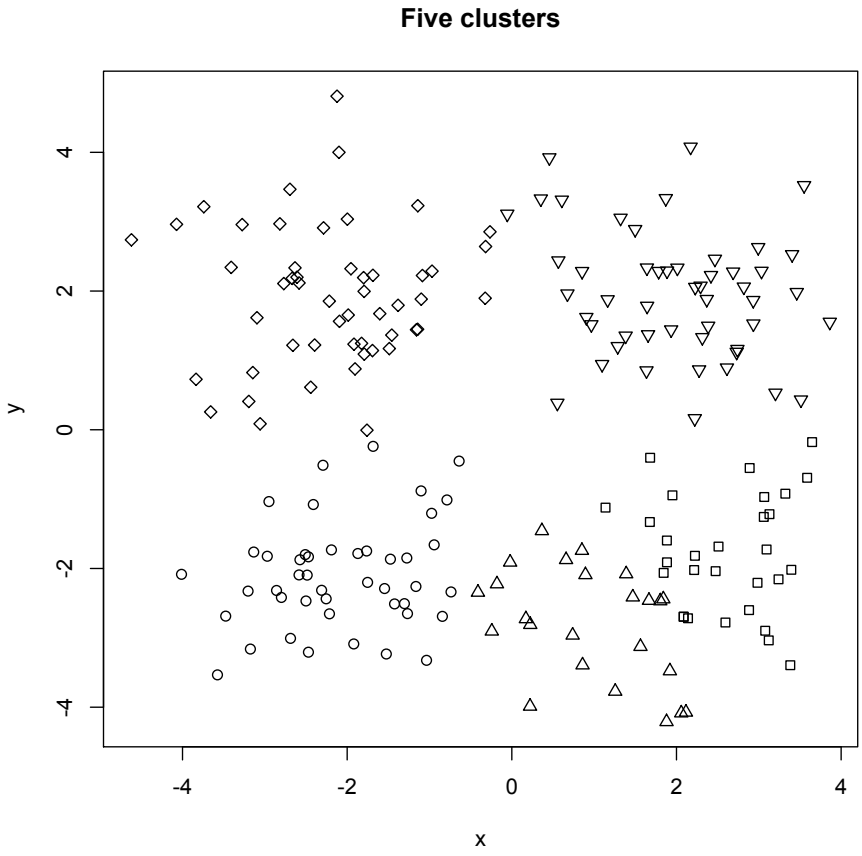
   (c) (5 points) Explain what the sum-of-squares is for k-means.

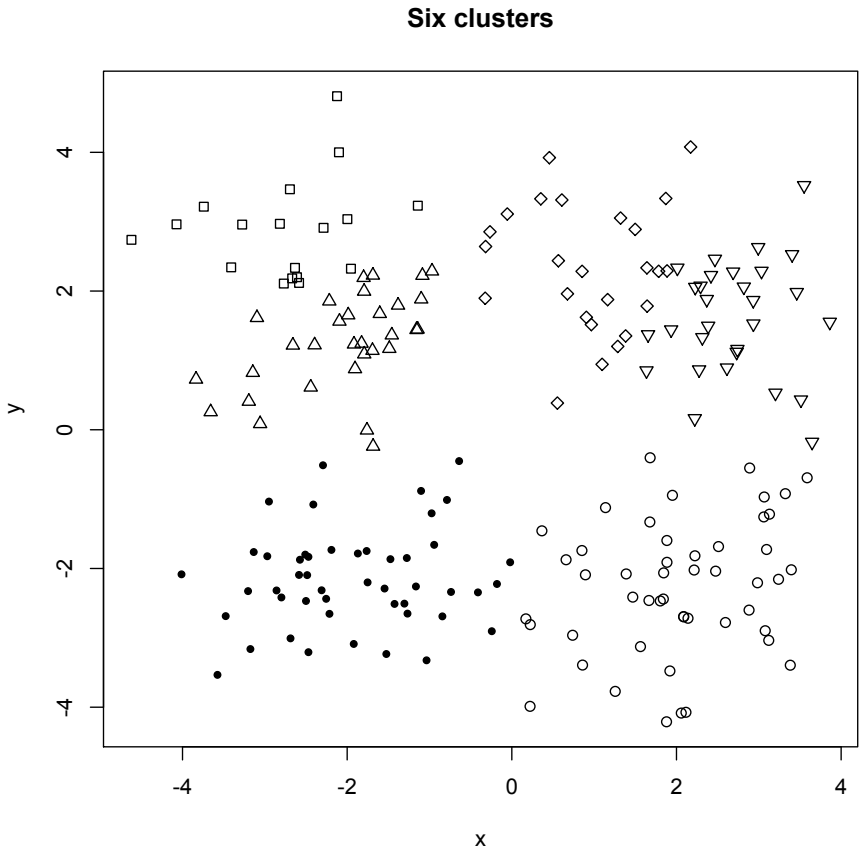
   (d) (8 points) The next pages show the results of clustering the same data with k means, with k running from 2 to 6; also a plot of the sum-of-squares versus k. How many clusters would you guess this data has, and why? Does it matter whether the plot is an average over many runs of the algorithm?

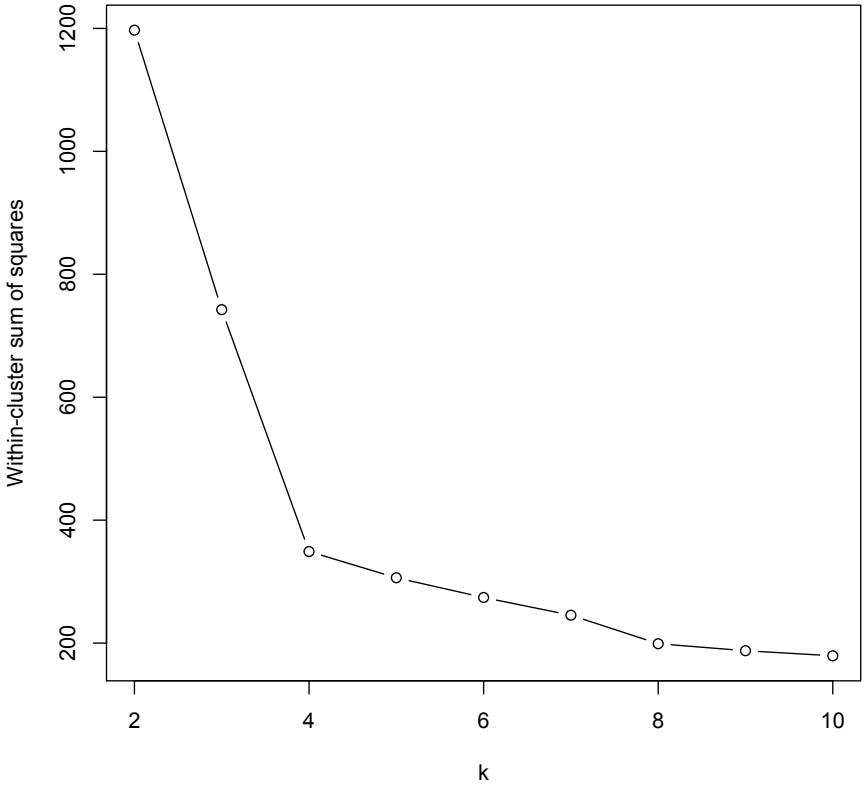

Two clusters
y
x
4
2
0
-2
-4
-4
-2
0
2
4


Three clusters
y
x
4
2
0
-2
-4
-4
-2
0
2
4

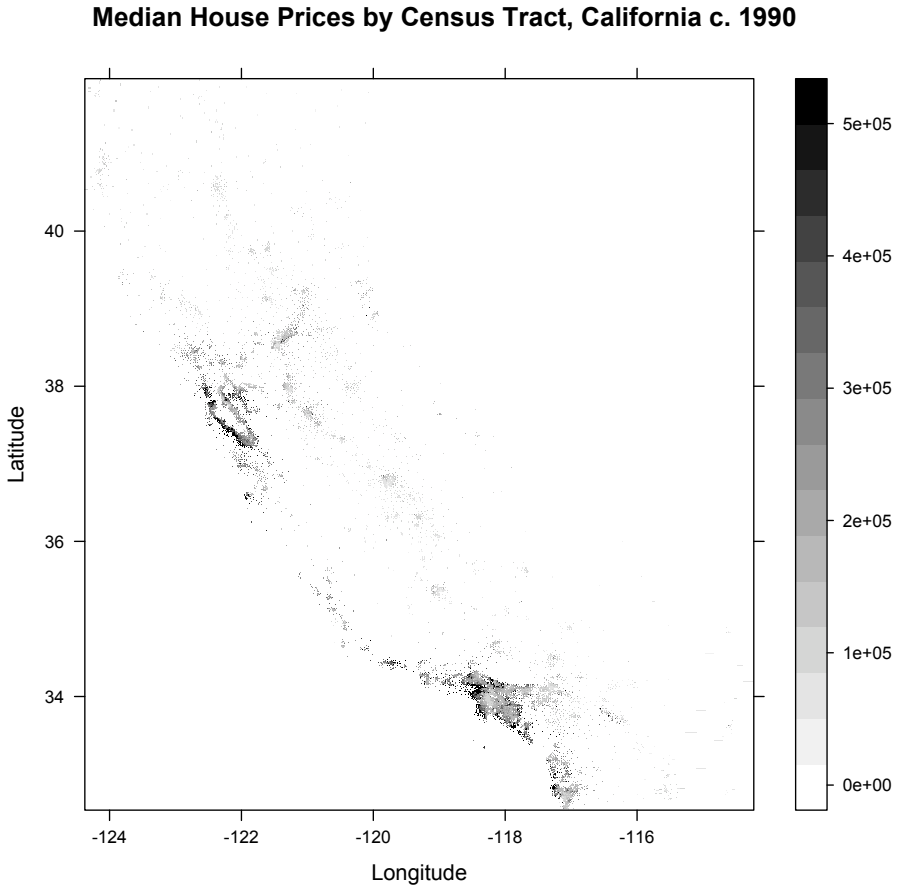

Four clusters
y
x
4
2
0
-2
-4
-4
-2
0
2
4


Five clusters
y
x
4
2
0
-2
-4
-4
-2
0
2
4


Six clusters
y
x
4
2
0
-2
-4
-4
-2
0
2
4


Within-cluster sum of squares
200
400
600
800
1000
1200
2
4
6
8
10
k

3. *What's That Got to Do with the Price of Houses in California?* The Census Bureau divides the country up into "tracts" of approximately equal population. For the 1990 Census, California was divided into 20640 tracts. One of the standard data sets (`housing` on `lib.stat.cmu.edu`) records the following for each tract in California: Median house price, median house age, average number of rooms per house, average number of bedrooms, average number of occupants, total number of houses, median income (in thousands of dollars), latitude and longitude.

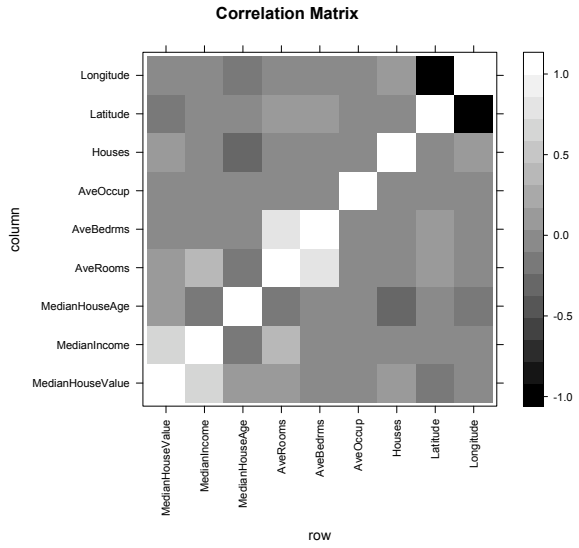


(a) (6 points) Here is the correlation matrix for the features.

	MedianHouseValue	MedianIncome	MedianHouseAge	AveRooms	AveBedrms
MedianHouseValue	1.00	0.69	0.11	0.15	-0.05
MedianIncome	0.69	1.00	-0.12	0.33	-0.06
MedianHouseAge	0.11	-0.12	1.00	-0.15	-0.08
AveRooms	0.15	0.33	-0.15	1.00	0.85
AveBedrms	-0.05	-0.06	-0.08	0.85	1.00
AveOccup	-0.02	0.02	0.01	0.00	-0.01
Houses	0.07	0.01	-0.30	-0.08	-0.05
Latitude	-0.14	-0.08	0.01	0.11	0.07
Longitude	-0.05	-0.02	-0.11	-0.03	0.01

	AveOccup	Houses	Latitude	Longitude
MedianHouseValue	-0.02	0.07	-0.14	-0.05
MedianIncome	0.02	0.01	-0.08	-0.02
MedianHouseAge	0.01	-0.30	0.01	-0.11
AveRooms	0.00	-0.08	0.11	-0.03
AveBedrms	-0.01	-0.05	0.07	0.01
AveOccup	1.00	-0.03	0.00	0.00
Houses	-0.03	1.00	-0.07	0.06
Latitude	0.00	-0.07	1.00	-0.92
Longitude	0.00	0.06	-0.92	1.00

Which variables would you guess are most important in predicting the housing price? Does it seem like any of them can be dropped? (Below is a visual summary of the matrix which you may find helpful.)



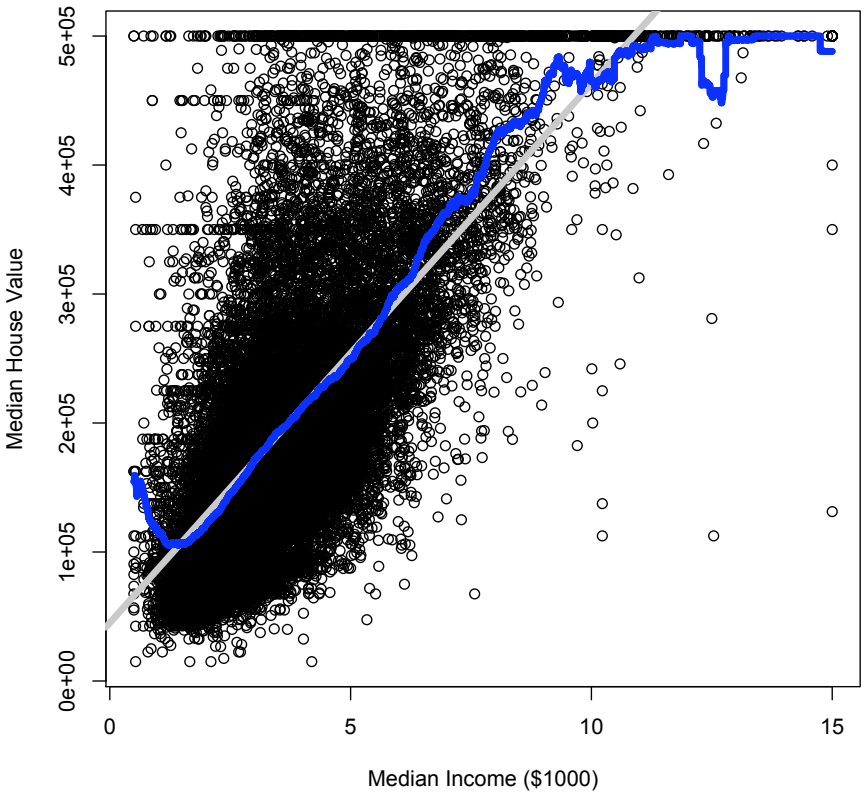
(b) (8 points) The next figure shows a plot of median house value versus median income, together with two fits. The straight line is the least-squares linear regression,

```
lm(CaliforniaHousing$MedianHouseValue ~ CaliforniaHousing$MedianIncome)
```

and the wiggly line is the kernel regression,

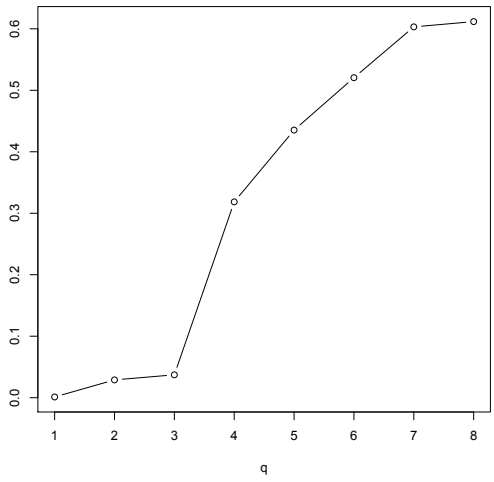
```
ksmooth(CaliforniaHousing$MedianIncome,CaliforniaHousing$MedianHouseValue)
```

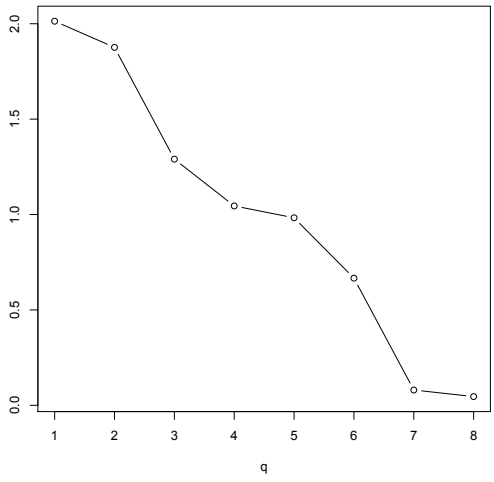
The R^2 of the linear model is 0.47 and that of the kernel-regression is 0.51. Which model do you prefer, and why? Should we consider adding more variables?



(c) What is the following code trying to do? (6 points) Is this equivalent to making a scree plot? (1 point) Which of the two figures following do you think it produced? (1 point)

```
> pca.of.predictors = prcomp(CaliforniaHousing[,2:9],scale.=TRUE)
> r2.with.q = function(q) {
  fit = lm(CaliforniaHousing$MedianHouseValue ~ pca.of.predictors$x[,1:q])
  return(summary(fit)$r.squared)
}
> plot(sapply(1:8,r2.with.q),xlab="q",ylab="",type="b")
```



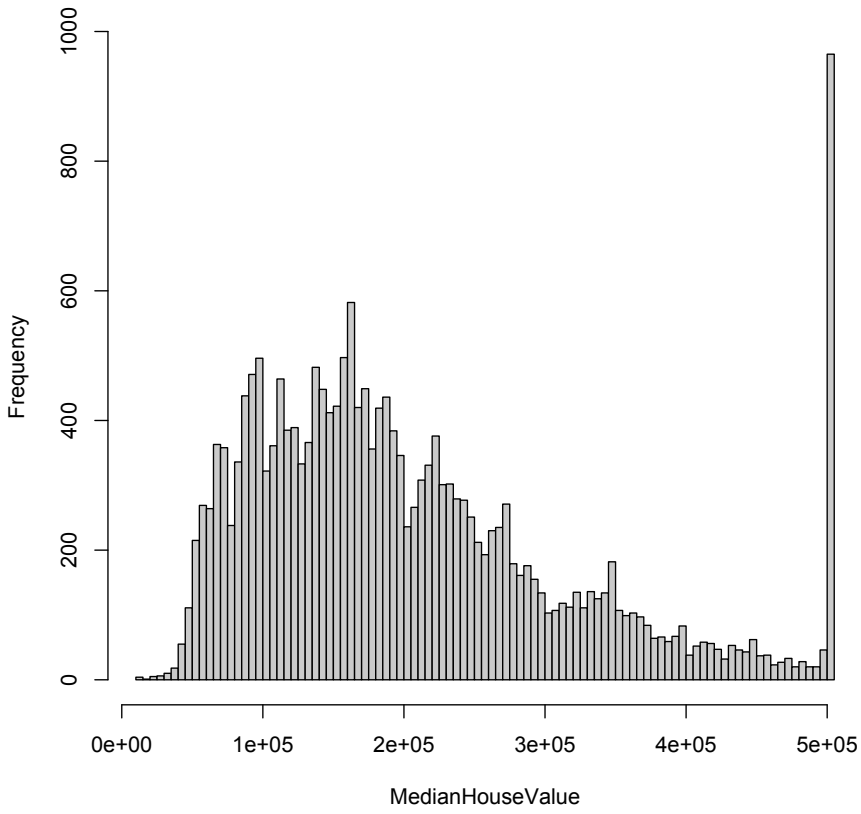


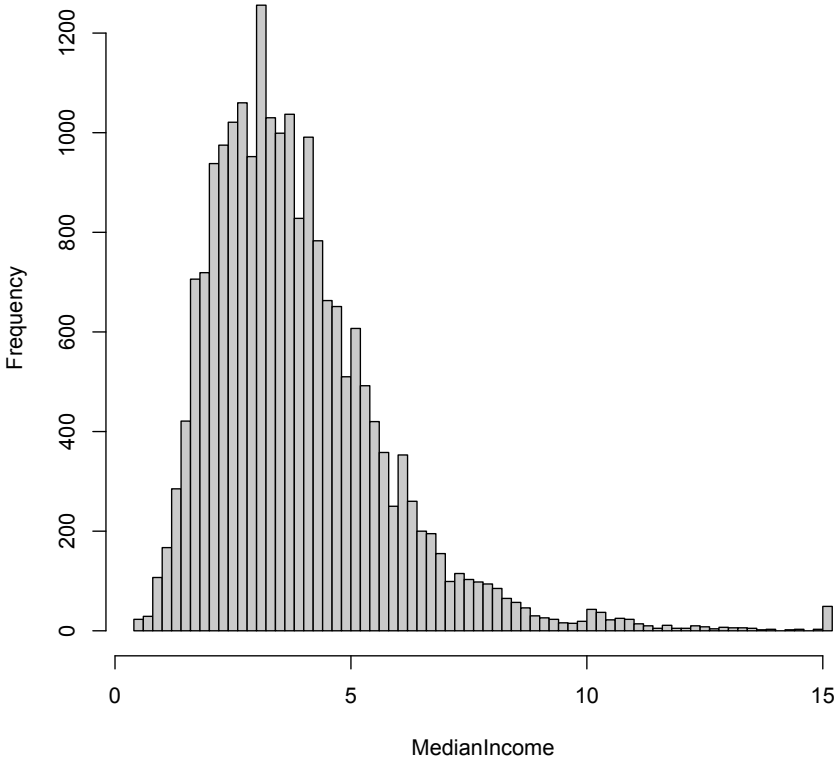
(d) (6 points) Consider the following R output.

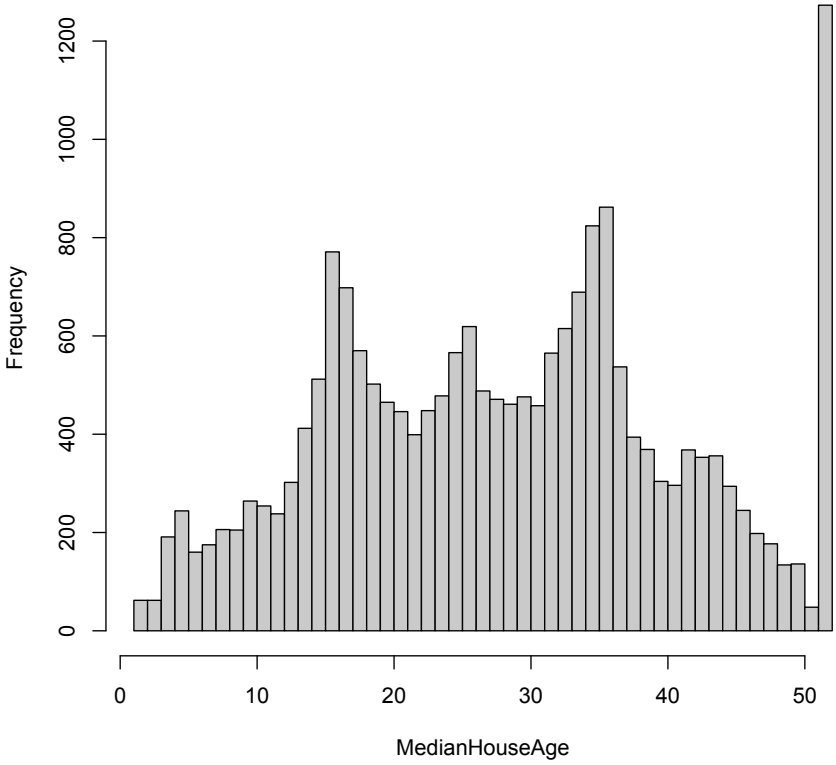
```
> summary(CaliforniaHousing)
 MedianHouseValue  MedianIncome  MedianHouseAge    AveRooms        AveBedrms
 Min.   : 14999   Min.   : 0.5   Min.   : 1     Min.   :  0.85   Min.   : 0.33
 1st Qu.:119600   1st Qu.: 2.6   1st Qu.:18     1st Qu.:  4.44   1st Qu.: 1.01
 Median :179700   Median : 3.5   Median :29     Median :  5.23   Median : 1.05
 Mean   :206856   Mean   : 3.9   Mean   :29     Mean   :  5.43   Mean   : 1.10
 3rd Qu.:264725   3rd Qu.: 4.7   3rd Qu.:37     3rd Qu.:  6.05   3rd Qu.: 1.10
 Max.   :500001   Max.   :15.0   Max.   :52     Max.   :141.91   Max.   :34.07
    AveOccup           Houses        Latitude    Longitude
 Min.   :   0.69   Min.   :   1   Min.   :33   Min.   :-124
 1st Qu.:   2.43   1st Qu.: 280   1st Qu.:34   1st Qu.:-122
 Median :   2.82   Median : 409   Median :34   Median :-118
 Mean   :   3.07   Mean   : 500   Mean   :36   Mean   :-120
 3rd Qu.:   3.28   3rd Qu.: 605   3rd Qu.:38   3rd Qu.:-118
 Max.   :1243.33   Max.   :6082   Max.   :42   Max.   :-114
```

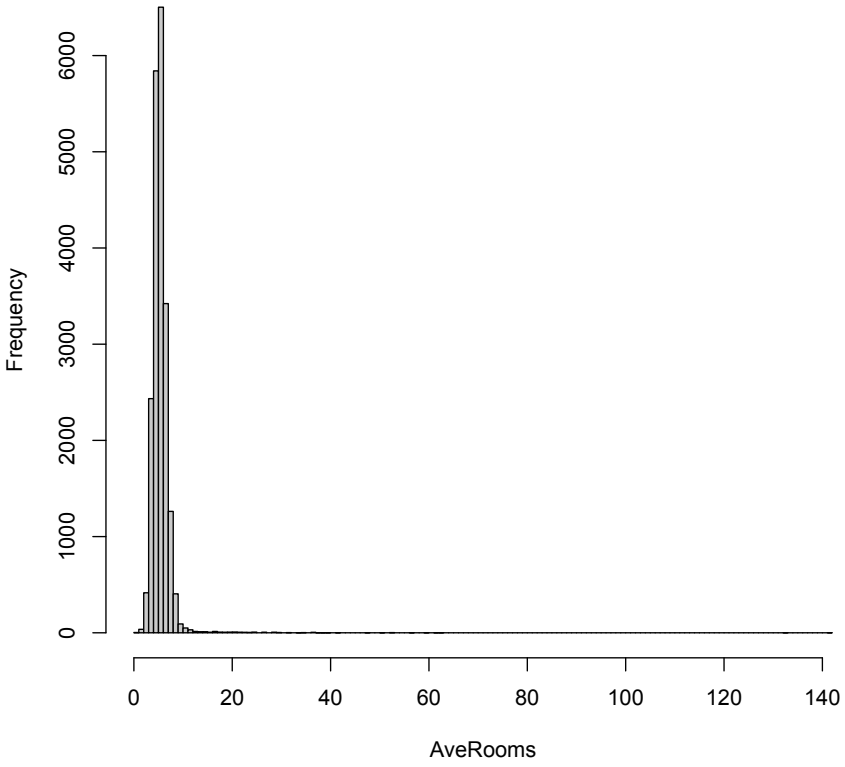
What is this command supposed to do? Does anything seems noteworthy about the output?

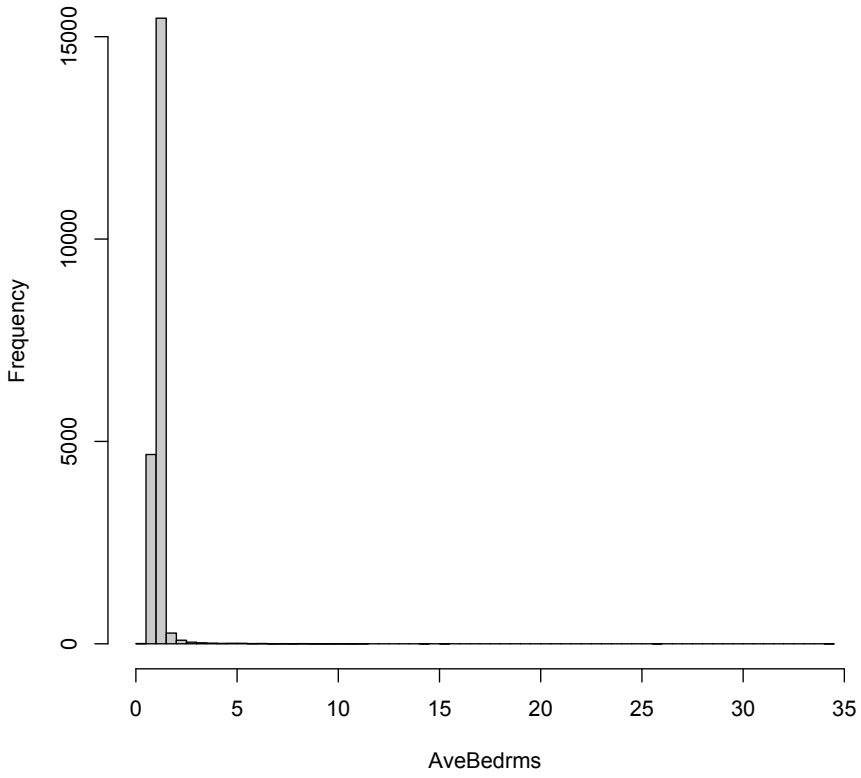
(e) (6 points) Looking at the output in the last command, and the histograms of the features on the next pages, does anything strike you about the features? Does any of it seem relevant to any of the previous questions?

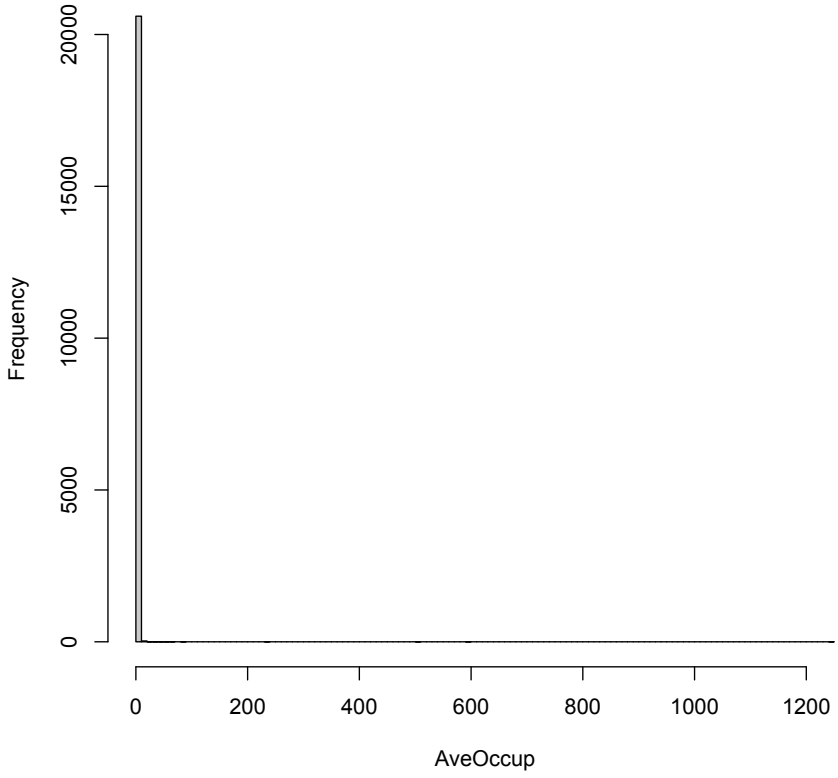

Frequency
0
200
400
600
800
1000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
MedianHouseValue

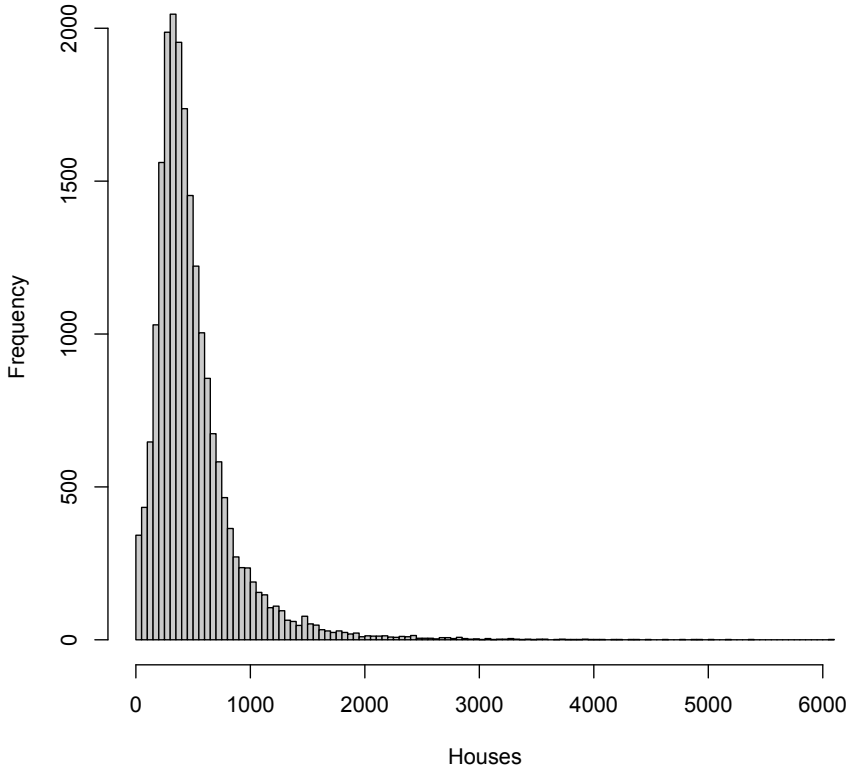

Frequency
0
200
400
600
800
1000
1200
0
5
10
15
MedianIncome

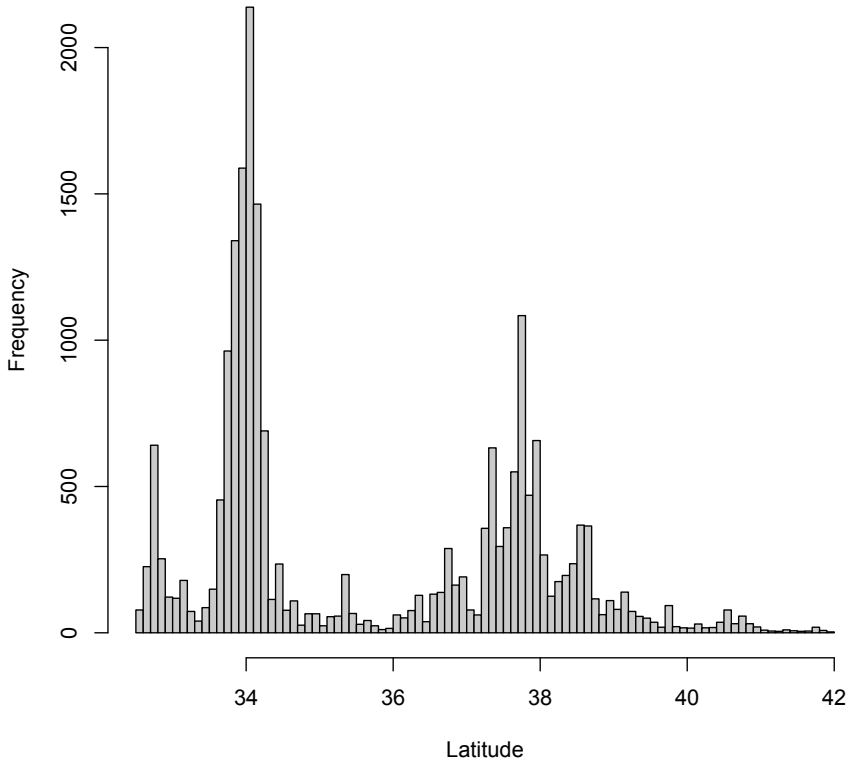

Frequency
0
200
400
600
800
1000
1200
0
10
20
30
40
50
MedianHouseAge

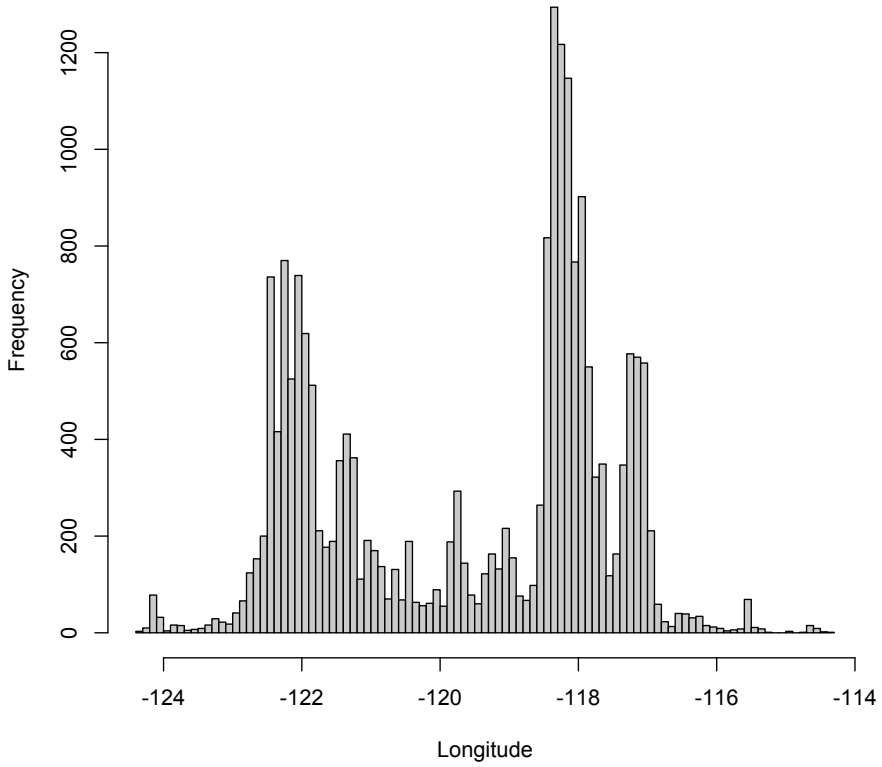

Frequency
0
1000
2000
3000
4000
5000
6000
0
20
40
60
80
100
120
140
AveRooms


Frequency
0
5000
10000
15000
0
5
10
15
20
25
30
35
AveBedrms


Frequency
0
5000
10000
15000
20000
0
200
400
600
800
1000
1200
AveOccup


Frequency
0
500
1000
1500
2000
0
1000
2000
3000
4000
5000
6000
Houses


Frequency
0
500
1000
1500
2000
34
36
38
40
42
Latitude


Frequency
0
200
400
600
800
1000
1200
-124
-122
-120
-118
-116
-114
Longitude