101chw2

Lucy Shao304575686

April 19, 2017

Q1

Make sure you have ggplot2 installed in Rstudio. Ask me or the TA if you need help with this. You will then need to enter require(ggplot2). You only need to do this once (but will have to do it again if you quit Rstudio and return and want to run ggplot2). Use ggplot2 to create a graphic (trying to predict the house price), based on the LArealestate.csv data from week 3 that shows us 3 (or more) variables on the same plot. What questions does the graphic answer?

```
library(ggplot2)
```

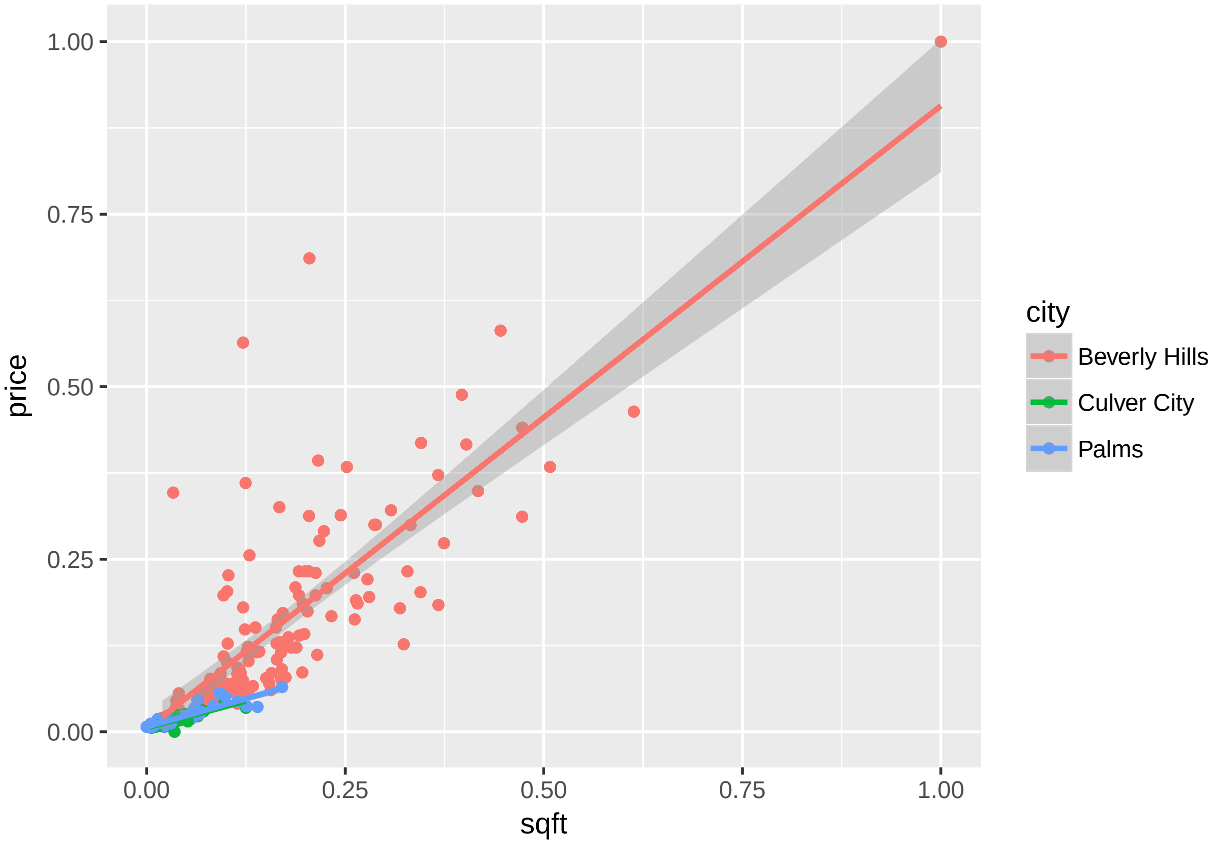
```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
data<-read.csv("/Users/lucy/Downloads/LArealestate.csv",header=TRUE,sep=",")
data<-data[complete.cases(data),]
attach(data)
norm<-function(x){return((x-min(x))/(max(x)-min(x)))}
data<-cbind(as.data.frame(lapply(data[,c(2:4,6)],norm)),city,type)
attach(data)
```

```
## The following objects are masked from data (pos = 3):
##
##     baths, beds, city, price, sqft, type
```

```
data[city=="culver city",]$city="Culver City"
qplot(sqft,price,color=city,data=data,geom=c("point","smooth"),method="lm")
```

```
## Warning: Ignoring unknown parameters: method
```



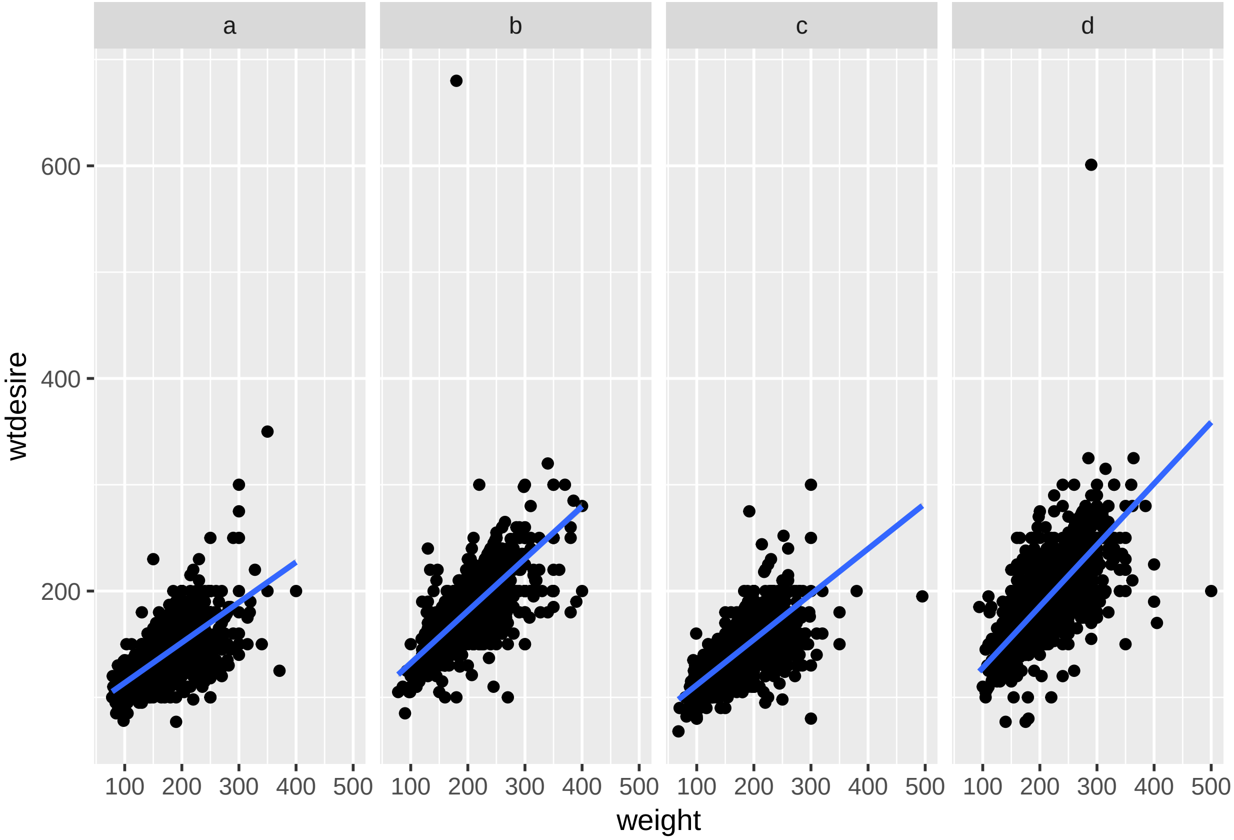
The graphic answers the question that how does the area of the house affects the price in different cities?

Q2

(a) Using the cdc.csv data (posted under Week 3) and ggplot2, make a plot that helps us understand the association between people's desired weight and their current weight, given their gender and whether or not they exercise. Your plot should include least squares lines to show the linear relation between desired weight and current weight for each of the four subgroups. Interpret these plots. (b) Instead of a regression line, use a smoother. Explain how the results differ from (a). Note: You can learn more about it at http://www.cdc.gov/brfss. The data come from the Behavioral Risk Factor Survey System.

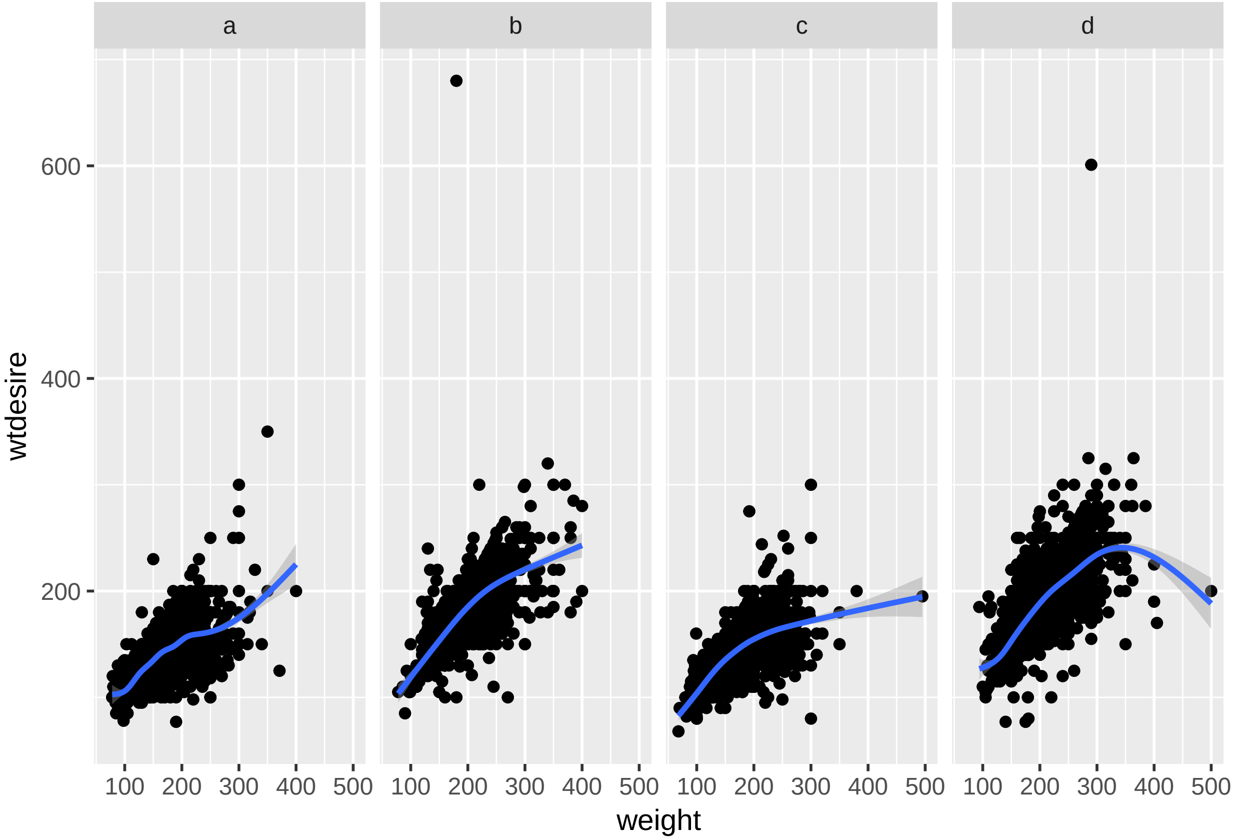
```
data2<-read.csv("/Users/lucy/Downloads/cdc.csv",header=TRUE)
data2$exerany<-as.factor(data2$exerany)
data2$type[data2$exerany=="0"&data2$gender=="f"]<-"a"
data2$type[data2$exerany=="0"&data2$gender=="m"]<-"b"
data2$type[data2$exerany=="1"&data2$gender=="f"]<-"c"
data2$type[data2$exerany=="1"&data2$gender=="m"]<-"d"
#a
qplot(weight,wtdesire,data=data2,facets = .~type,geom = c("point","smooth"),method="lm")
```

```
## Warning: Ignoring unknown parameters: method
```



```
#b
qplot(weight,wtdesire,data=data2,facets = .~type,geom = c("point","smooth"))
```

```
## `geom_smooth()` using method = 'gam'
```



```
#it fits the data more accuratly, but it does not show a linear relationship which can be used easily f
```

Q3

Banknote data posted on week 3: This Data is on real and counterfeit banknotes Y = 1 means counterfeit and Y = 0 means real. The goal is to use the knn classification algorithm to classify bank notes based on the following features. i.e. Develop a rule to tell them apart.

```
data3<-read.table("/Users/lucy/Downloads/banknote.txt",header = T,sep="\t")
attach(data3)
bank<-cbind(as.data.frame(lapply(data3[,c(1:6)],norm)),Y)
#summary(bank)
bank$Y<-as.factor(bank$Y)
#split
set.seed(33445566)
sample1<-sample(seq(1,200),140,replace=F)
train<-bank[sample1,]
test<-bank[-sample1,]
library(class)
##k=1
m<-knn(train=train[,1:6],test=test[,1:6],cl=train[,7],k=1)
table(m)
```

```
## m
##  0  1
## 33 27
```

```r
table<-table(test[,7],m)
table
```

```
##    m
##      0  1
##   0 33  1
##   1  0 26
```

```r
accuracy<-(table[1,1]+table[2,2])/sum(table)
accuracy
```

```
## [1] 0.9833333
```

```r
#k=3
m3<-knn(train=train[,1:6],test=test[,1:6],cl=train[,7],k=3)
table(m3)
```

```
## m3
##  0  1
## 34 26
```

```r
table3<-table(test[,7],m3)
table3
```

```
##    m3
##      0  1
##   0 34  0
##   1  0 26
```

```r
accuracy3<-(table3[1,1]+table3[2,2])/sum(table3)
accuracy3
```

```
## [1] 1
```

```r
#k=5
m5<-knn(train=train[,1:6],test=test[,1:6],cl=train[,7],k=5)
table(m5)
```

```
## m5
##  0  1
## 33 27
```

```r
table5<-table(test[,7],m5)
table5
```

```
##    m5
##      0  1
##   0 33  1
##   1  0 26
```

```r
accuracy5<-(table5[1,1]+table5[2,2])/sum(table5)
accuracy5
```

```
## [1] 0.9833333
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.3.2
```

```
## Loading required package: lattice
```

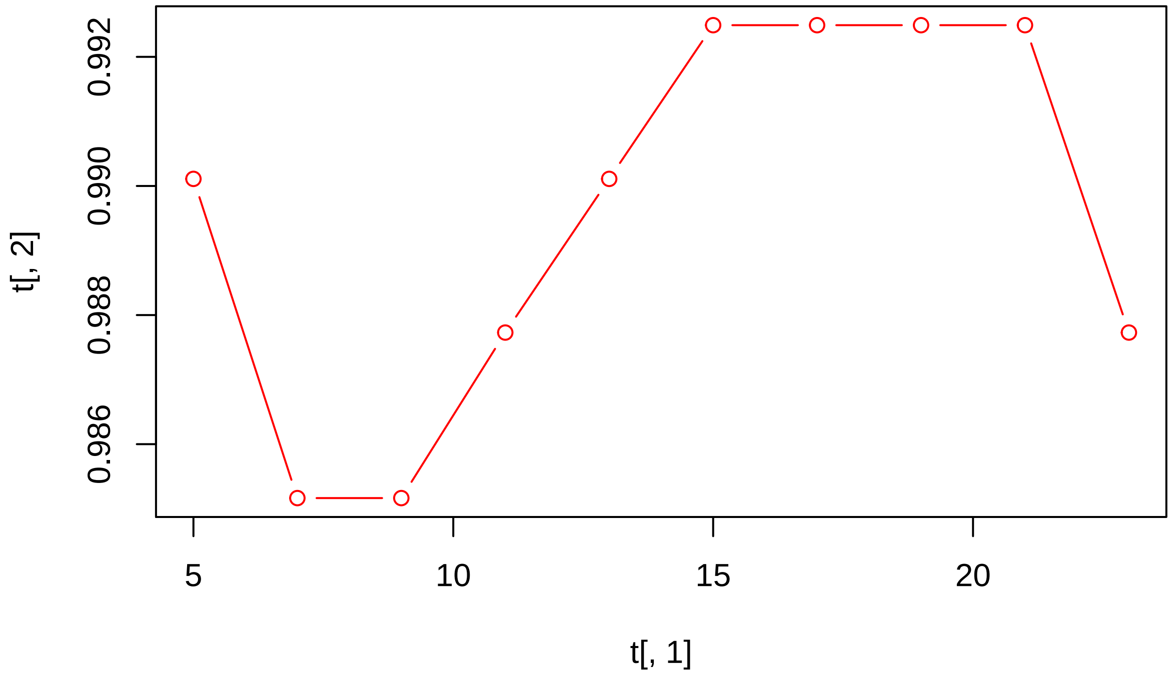
```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.3.2
```

```r
a<-trainControl(method="repeatedcv",number=10,repeats = 3)
set.seed(33445566)
k<-train(Y~.,data=train,method="knn",trControl=a, preProcess=c("center","scale"),tuneLength=10)
k
```

```
## k-Nearest Neighbors
##
## 140 samples
##   6 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (6), scaled (6)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 126, 125, 126, 126, 127, 126, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    5  0.9901099  0.9800344
##    7  0.9851648  0.9701745
##    9  0.9851648  0.9701745
##   11  0.9877289  0.9752725
##   13  0.9901099  0.9800344
##   15  0.9924908  0.9847963
##   17  0.9924908  0.9847963
##   19  0.9924908  0.9847963
##   21  0.9924908  0.9847963
##   23  0.9877289  0.9752725
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was k = 21.
```

```r
t<-data.frame(k[4])
plot(t[,1],t[,2],type="b",col="red")
```



```
#k=15,17,19,21,
```

Q4) Question 2.4.7 abc

X1 X2 X3 Distance Y
1 0 3 0 3 Red
2 2 0 0 2 Red
3 0 1 3 3.2 Red
4 0 1 2 2.2 Green
5 -1 0 1 1.4 Green
6 1 1 1 1.7 Red

a)

(b) it is Green since we want the smallest distance so #5 is the closest neighbor for K = 1.

(c) it is Red. minimize the distance so Observations #2,#5,and# 6 are the closest neighbors for K = 3, where 2 is Red, 5 is Green, and 6 is Red.