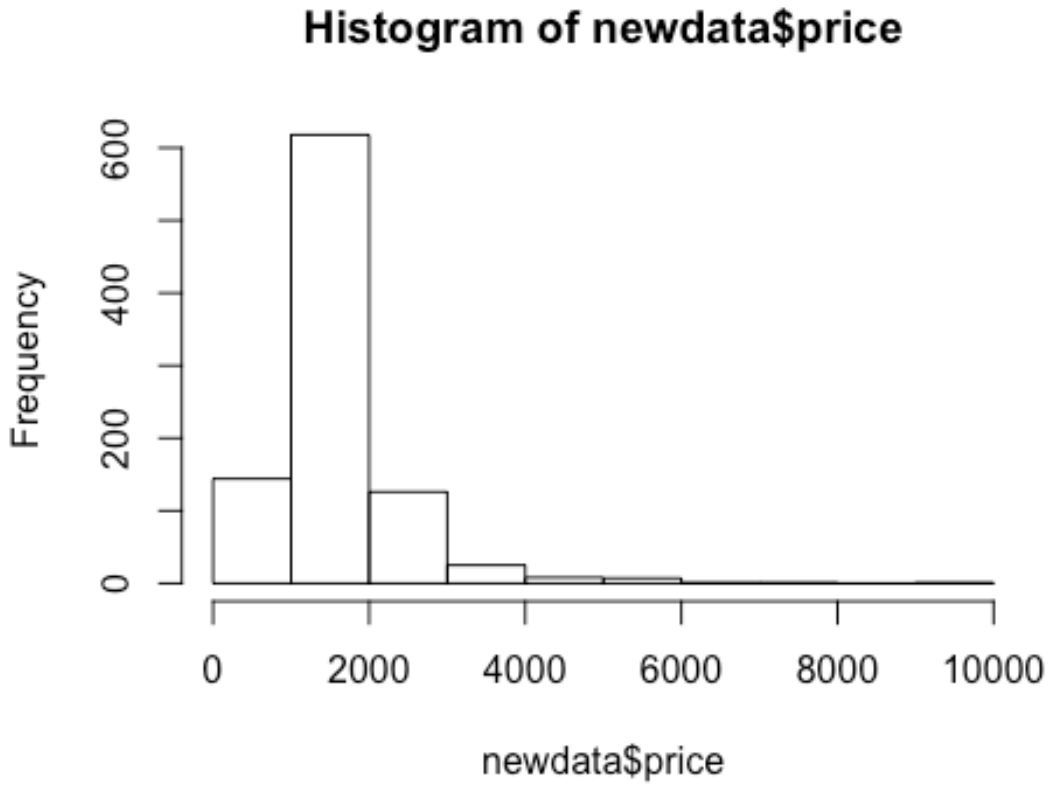
regression_rent

Preparing the regression

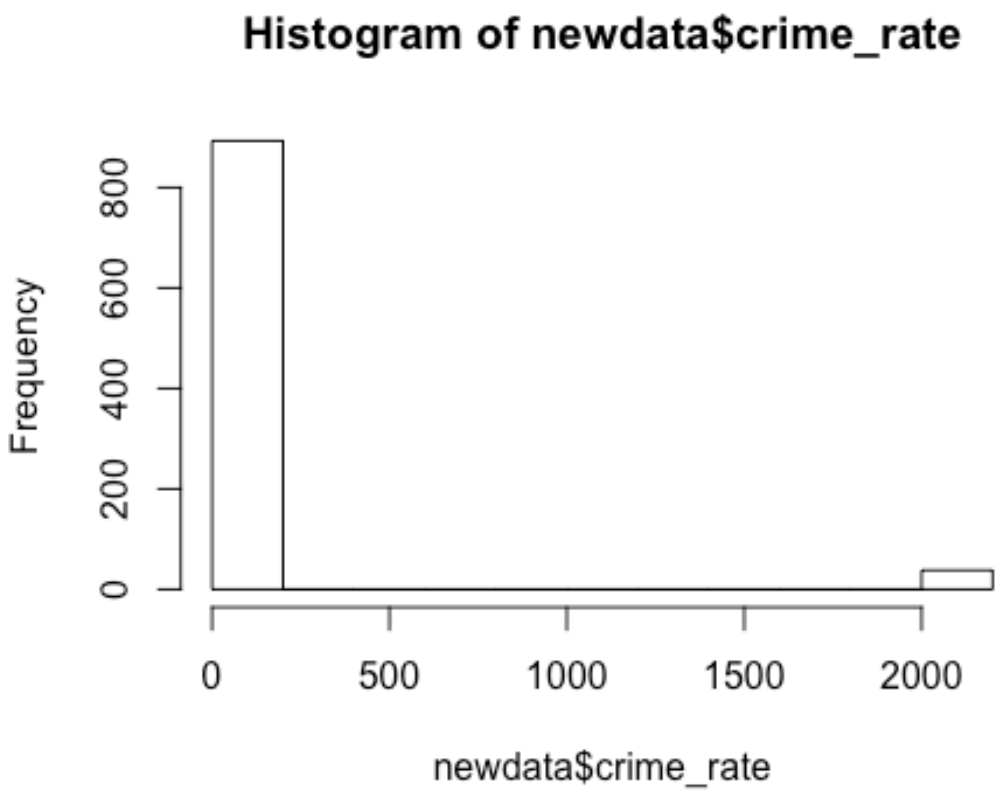
Since the sales data doesnt have the unit type so in the final model we take it out.

Changing the numeric variables to log form, because there are not normaly distributed.

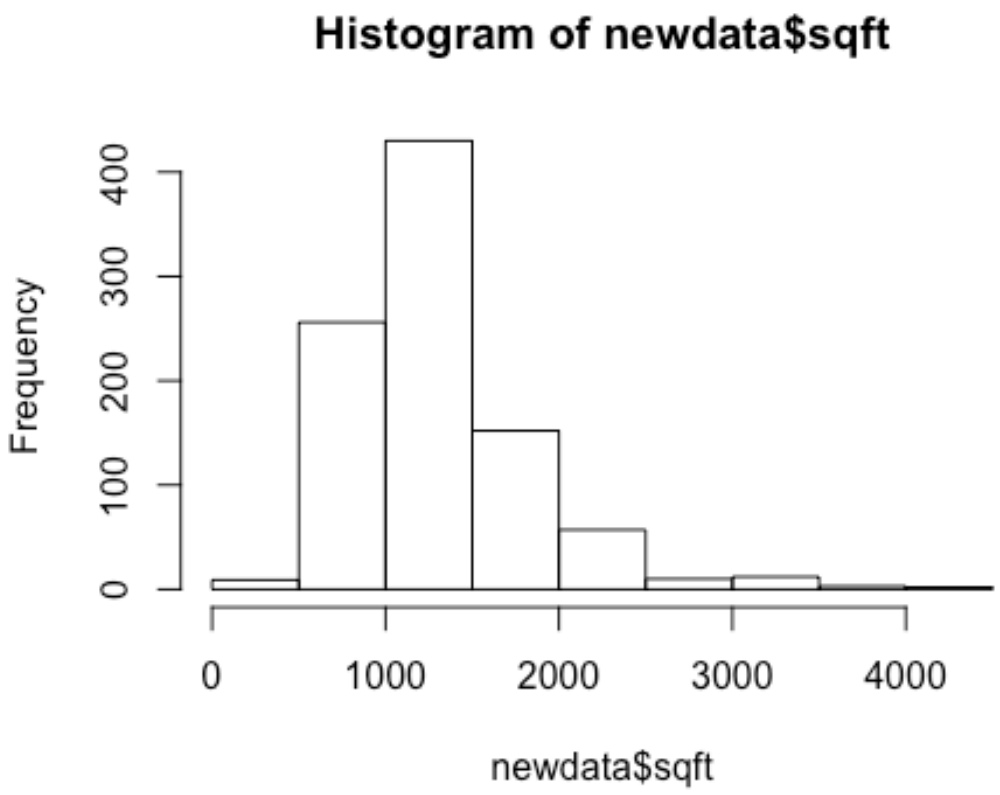
```
rent<-read.csv('rent.csv', header=TRUE)
rent$X <- NULL
rent$unittype<-NULL
newdata <- na.omit(rent)
hist(newdata$price)
```




```
hist(newdata$crime_rate)
```




```
hist(newdata$sqft)
```




```
newdata$rent <- log(newdata$price)
newdata$crime_rate<-log(newdata$crime_rate)
newdata$sqft<-log(newdata$sqft)
```

Right now, creating the train and test data set.

```
library(MASS)
ind<-sample(2,nrow(newdata),replace=TRUE,prob = c(0.8,0.2))
train<-newdata[ind==1,]
test<-newdata[ind==2,]
fit <- lm(rent~bedrooms+baths+distance+sqft+crime_rate,data=train)
step <- stepAIC(fit, direction="both")
```

```
## Start:  AIC=-1964.37
## rent ~ bedrooms + baths + distance + sqft + crime_rate
##
##              Df Sum of Sq    RSS     AIC
## - bedrooms    1    0.0010 49.447 -1966.3
## <none>                    49.446 -1964.4
## - crime_rate  1    1.2561 50.702 -1948.0
## - distance    1    4.2024 53.648 -1906.6
## - sqft        1    7.0354 56.481 -1868.9
## - baths       1   10.9483 60.394 -1819.8
##
```

```
## Step:  AIC=-1966.35
## rent ~ baths + distance + sqft + crime_rate
##
##              Df Sum of Sq    RSS     AIC
## <none>                     49.447 -1966.3
## + bedrooms    1    0.0010 49.446 -1964.4
## - crime_rate  1    1.3178 50.764 -1949.1
## - distance    1    4.6838 54.130 -1902.0
## - sqft        1    8.2823 57.729 -1854.8
## - baths       1   11.2737 60.720 -1817.8

step$anova

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## rent ~ bedrooms + baths + distance + sqft + crime_rate
##
## Final Model:
## rent ~ baths + distance + sqft + crime_rate
##
##
##         Step Df     Deviance Resid. Df Resid. Dev       AIC
## 1                                  727   49.44557 -1964.368
## 2 - bedrooms  1 0.0009695676       728   49.44654 -1966.354

fiit<-lm(rent~bedrooms+baths+distance+sqft+crime_rate,data=train)
prediction<-predict(fiit,test)
prediction<-as.data.frame(prediction)
names(prediction) = c("pre")
prediction$pre<-as.numeric(prediction$pre)
D<-cbind(prediction,test) #combine the predicted value into the test dataset
D$Difference <- abs(D$rent - D$pre)
summary(D$rent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.515   7.090   7.313   7.370   7.576   8.698

summary(D$Difference)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0002827 0.0722100 0.1758000 0.2212000 0.3204000 1.0690000
```

We can see from above result, the bias between training and test set is not quiet obvious and this is the best model we can get. So, we decided to take it as our final model.

Using stepwise first to set up the final model and then get the statistic parameters.

```
fit <- lm(rent~bedrooms+baths+distance+sqft+crime_rate,data=newdata)
step <- stepAIC(fit, direction="both")

## Start:  AIC=-2447.46
## rent ~ bedrooms + baths + distance + sqft + crime_rate
##
##              Df Sum of Sq    RSS     AIC
## <none>                     66.322 -2447.5
## - bedrooms    1    0.1676 66.489 -2447.1
## - crime_rate  1    1.7645 68.086 -2425.0
## - distance    1    6.0514 72.373 -2368.2
## - sqft        1   10.5387 76.860 -2312.2
## - baths       1   13.5107 79.832 -2276.8

step$anova

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## rent ~ bedrooms + baths + distance + sqft + crime_rate
##
## Final Model:
## rent ~ bedrooms + baths + distance + sqft + crime_rate
##
##
##   Step Df Deviance Resid. Df Resid. Dev       AIC
## 1                        925   66.32177 -2447.461
```

Let's take a look of the statistic of our model

```
fitt<-lm(rent~baths+distance+sqft+crime_rate+bedrooms,data=newdata)
summary(fitt)

##
## Call:
## lm(formula = rent ~ baths + distance + sqft + crime_rate + bedrooms,
##     data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1991 -0.1723 -0.0076  0.1494  1.0388
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.497745   0.213146  21.102  < 2e-16 ***
## baths        0.207699   0.015130  13.727  < 2e-16 ***
## distance    -0.047339   0.005153  -9.187  < 2e-16 ***
## sqft         0.389812   0.032153  12.124  < 2e-16 ***
## crime_rate  -0.036110   0.007279  -4.961 8.35e-07 ***
## bedrooms    -0.018490   0.012096  -1.529    0.127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2678 on 925 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.5324
## F-statistic: 212.8 on 5 and 925 DF,  p-value: < 2.2e-16
```