Regression

Angus Huang

December 20, 2017

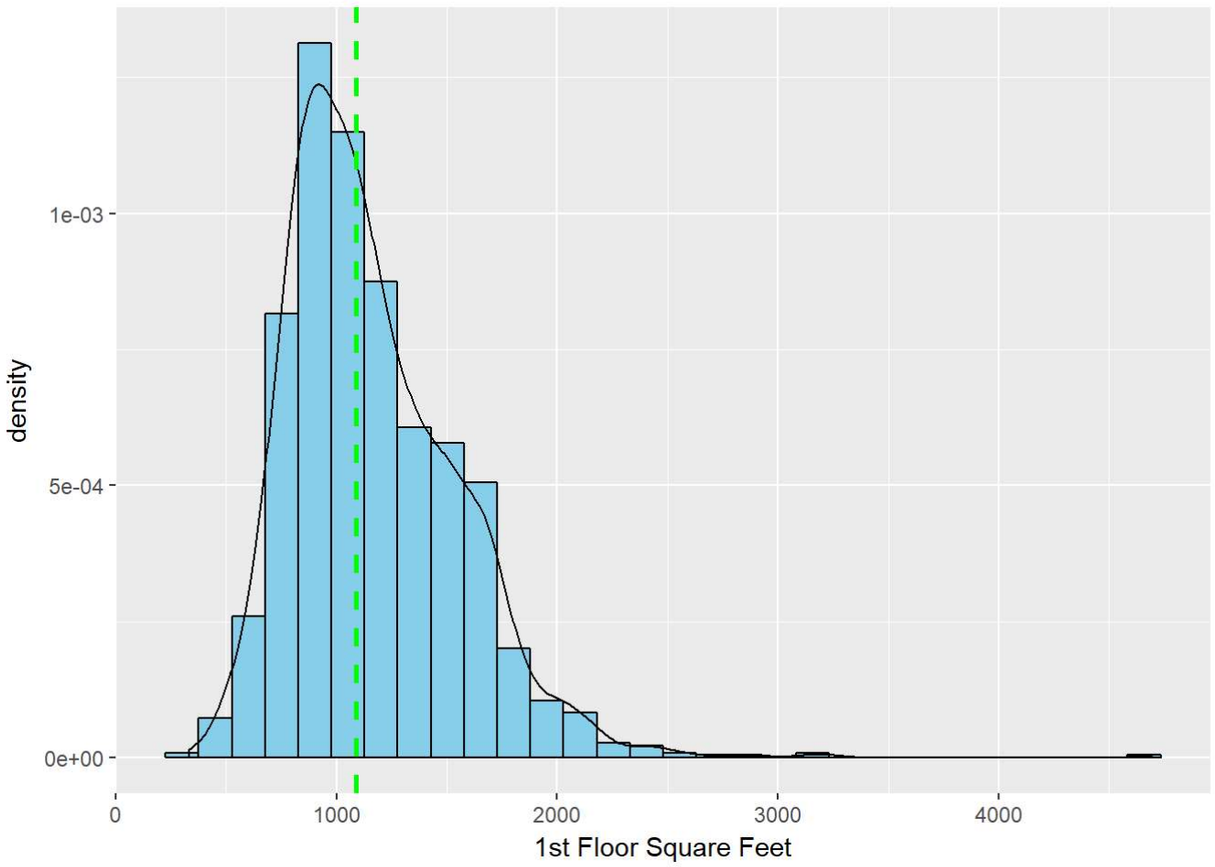
```
library(MASS)
library(ggplot2)
```

```
mydata <-read.table("https://github.com/angus001/Data605/raw/master/train.csv",header = T, sep=
",")
testdata <- read.table("https://github.com/angus001/Data605/raw/master/test.csv",header = T, sep
 = ",")
```

```
X<-mydata$X1stFlrSF
Y<-mydata$SalePrice
df<-data.frame(X,Y)
```

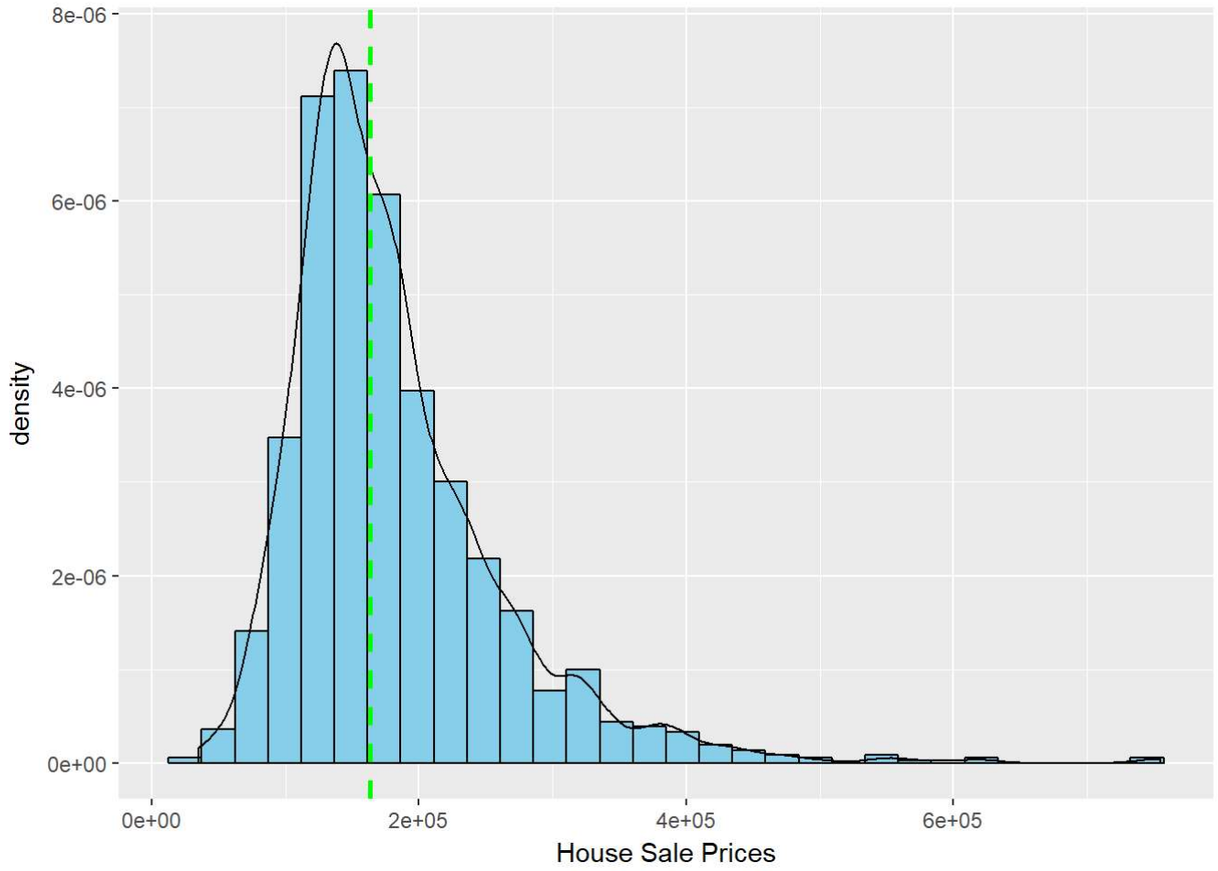
```
ggplot(df, aes(x=X)) +
  geom_histogram(aes(y=..density..), colour="black", fill="skyblue")+
  geom_density(alpha=0.5) +
  xlab("1st Floor Square Feet") +
  geom_vline(aes(xintercept=median(df$X)),
             color="green", linetype="dashed", size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



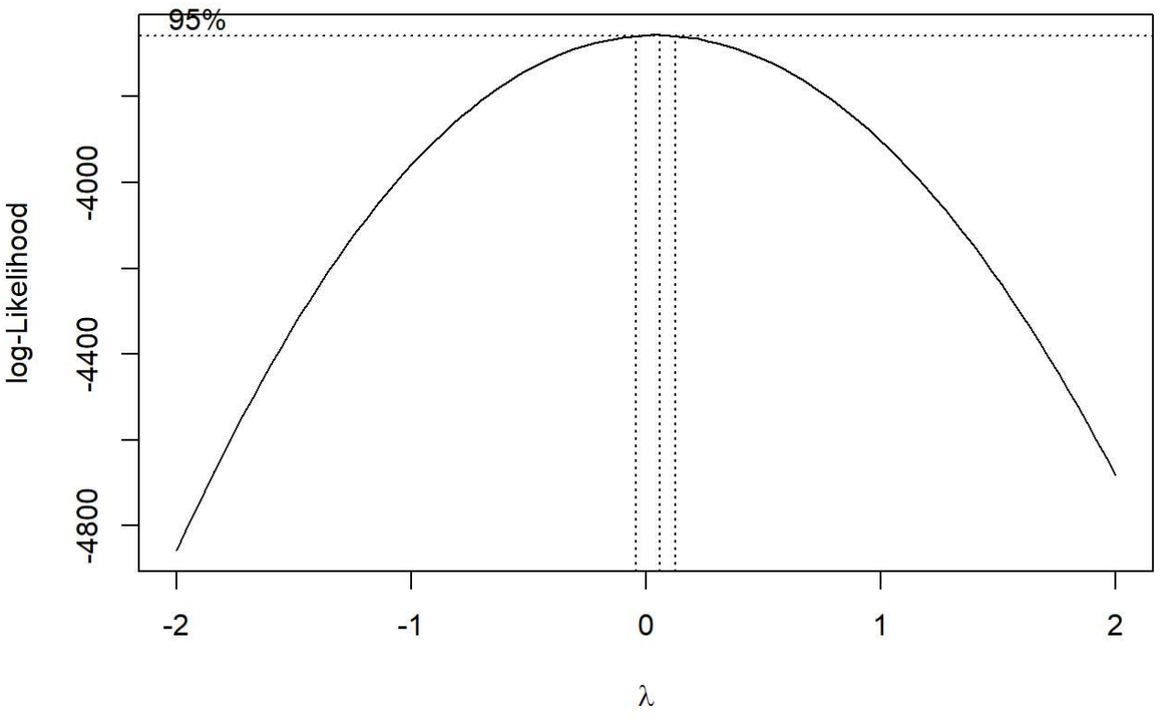

```
ggplot(df, aes(x=Y)) +
  geom_histogram(aes(y=..density..), colour="black", fill="skyblue")+
  geom_density(alpha=0.5) +
  xlab("House Sale Prices") +
  geom_vline(aes(xintercept=median(df$Y)),
             color="green", linetype="dashed", size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Perform boxcox analysis to find log-Likelihood. Look for Lambda value with max likelihood. THe max likelihood is at 0.06 power.

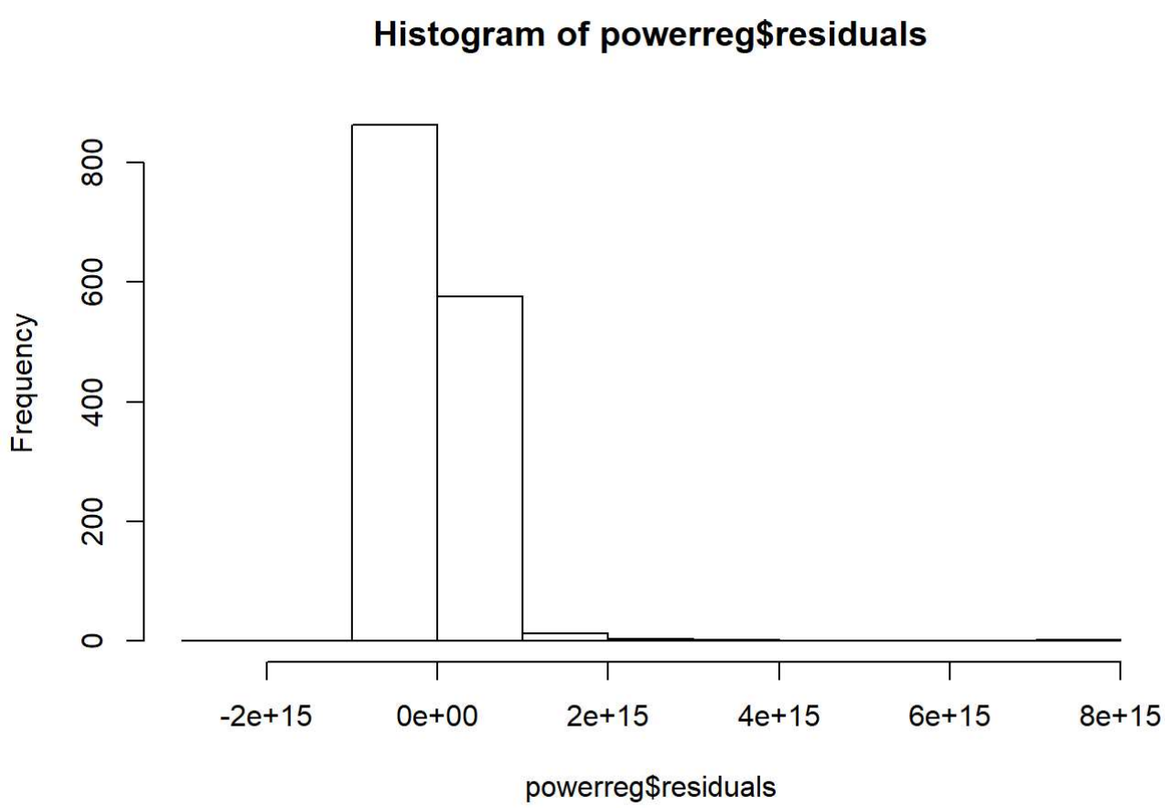
```
bc = boxcox(Y~X, data = df)
```



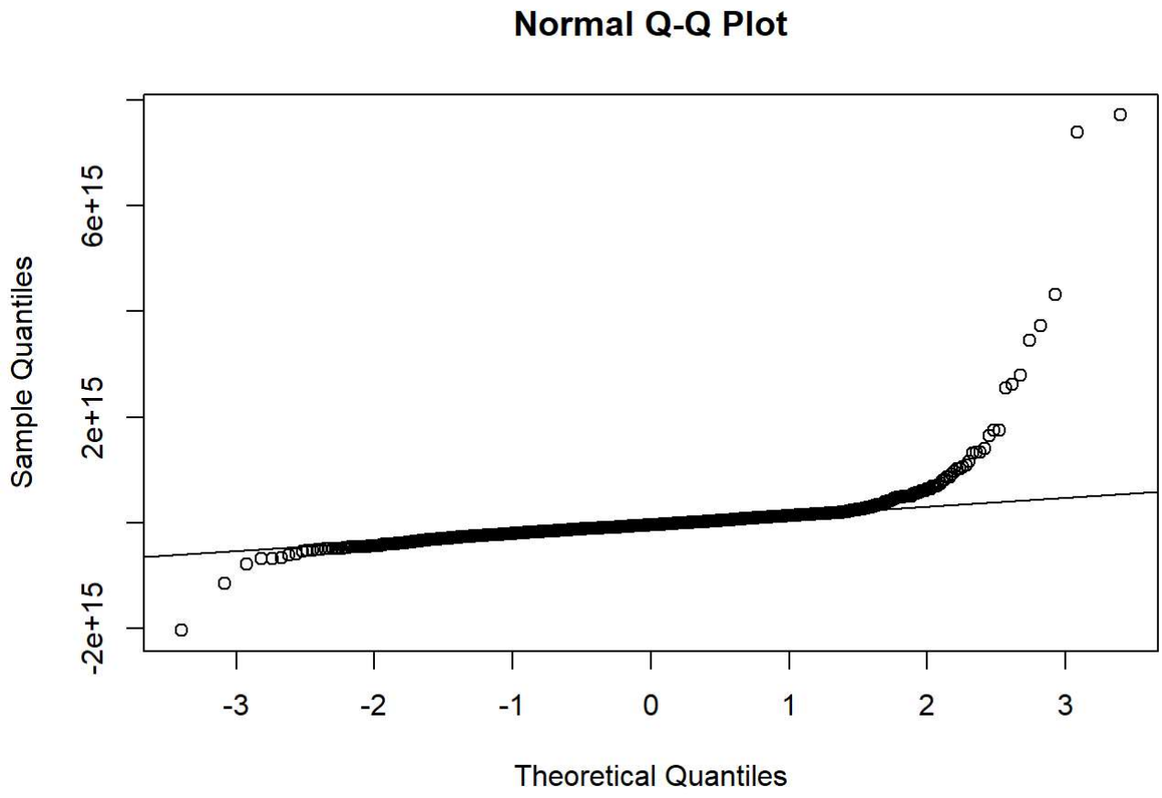

```
lamda =bc$x
likelihood = bc$y
bc1=cbind(lamda,likelihood)
head(bc1[order(-likelihood),])
```

```
##            lamda likelihood
## [1,]  0.06060606  -3656.618
## [2,]  0.02020202  -3656.622
## [3,]  0.10101010  -3657.490
## [4,] -0.02020202  -3657.505
## [5,]  0.14141414  -3659.237
## [6,] -0.06060606  -3659.270
```

```
df$Ypower = (df$Y)^3/50
powerreg <- lm(Ypower~X, df)
hist(powerreg$residuals)
```




```
qqnorm(powerreg$residuals)
qqline(powerreg$residuals)
```



Perform a correlation test between variables. The correlation test shows a correlation of 0.605 without tranforming the variable. The correlation actually become less to 0.441 after transforming the variable.

```
cor(df)
```

```
##                X         Y    Ypower
## X      1.0000000 0.6058522 0.4411002
## Y      0.6058522 1.0000000 0.8019417
## Ypower 0.4411002 0.8019417 1.0000000
```

```
cor.test(df$X,df$Y, conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$X and df$Y
## t = 29.078, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  0.5613896 0.6468270
## sample estimates:
##       cor
## 0.6058522
```

Fitting the data point into different distribution to understand the underlying spread of the data.

```
fit <- fitdistr(df$X, densfun = 'cauchy')
fit
```

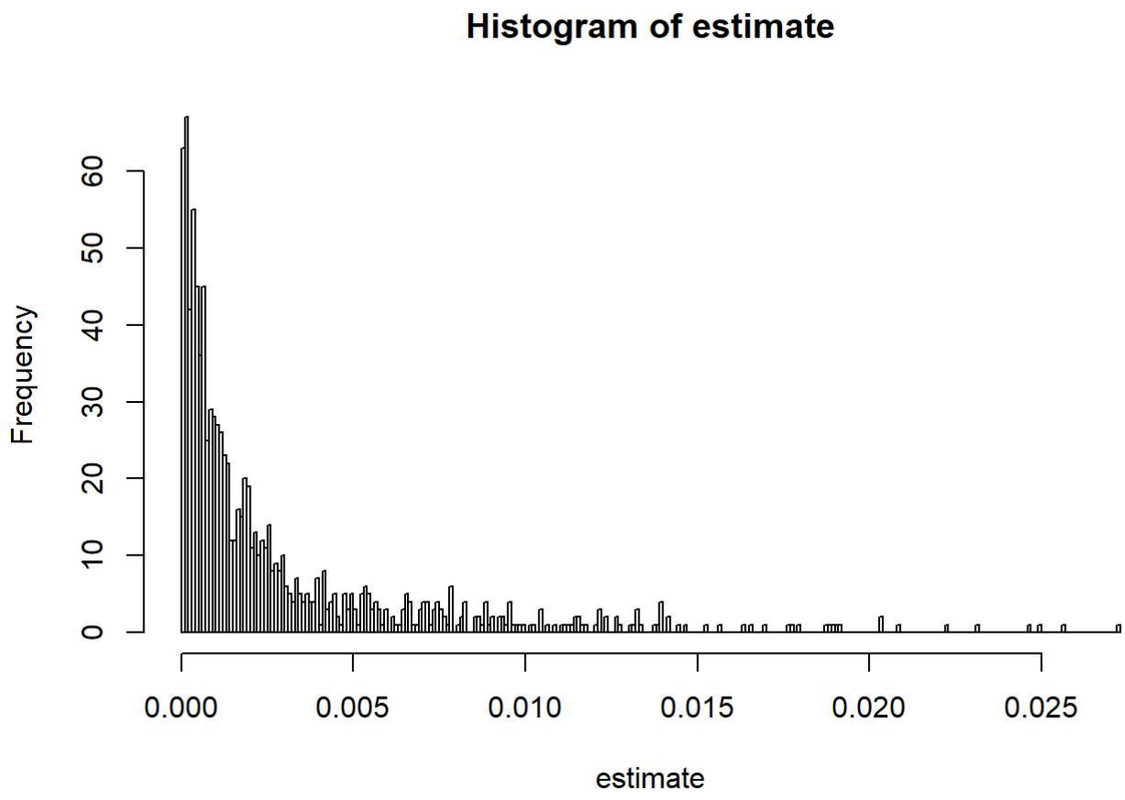
```
##     location        scale
##   1059.239655    212.473032
##  (   9.090705) (   7.210534)
```

```
lamda2 <- fit$estimate
lamda2
```

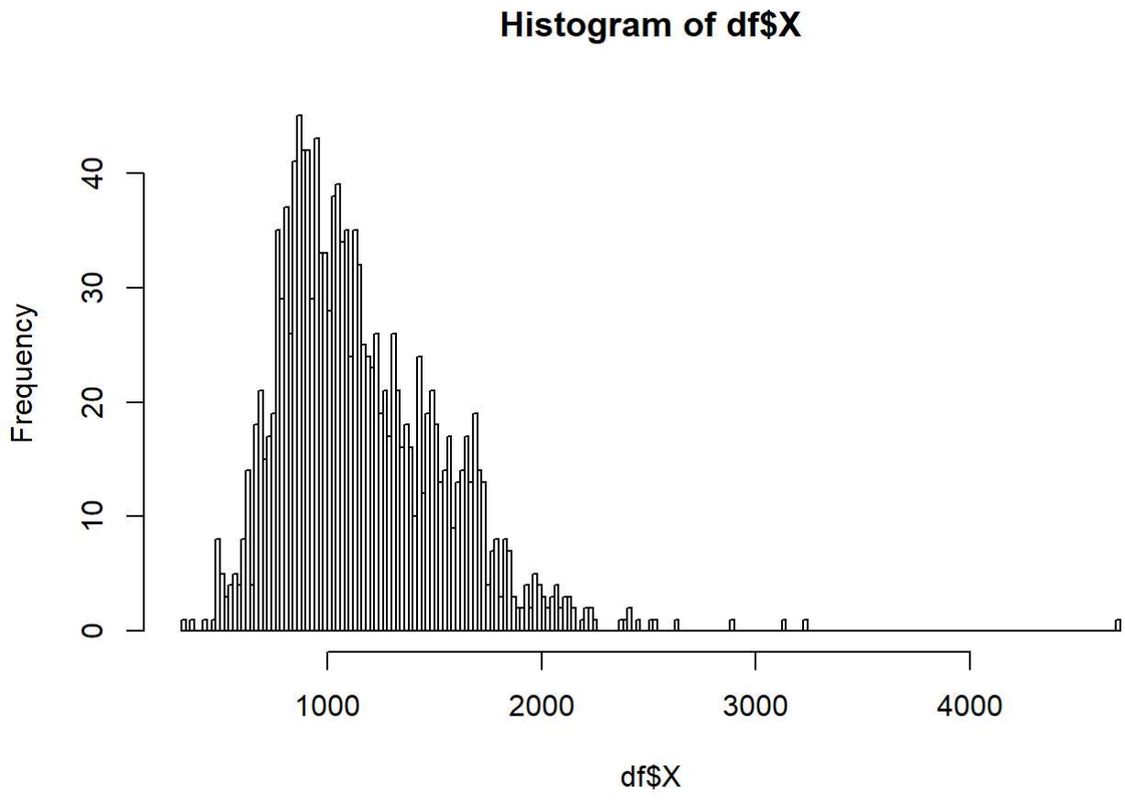
```
## location    scale
## 1059.240  212.473
```

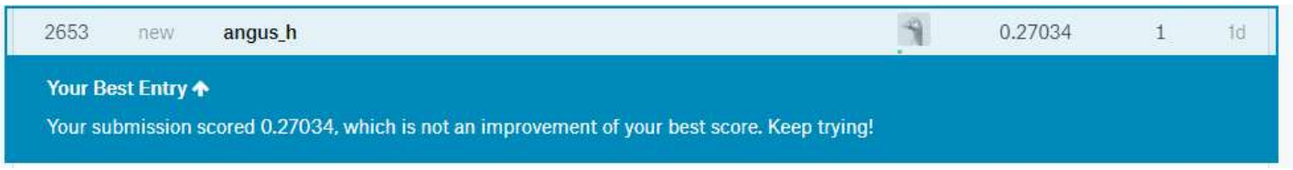
Take 1000 samples from the distribution, plot a histogram and compare with the non-transformed original values.

```
estimate <- rexp(1000,lamda2)
hist(estimate, breaks = 200)
```




```
hist(df$X, breaks = 200)
```





Kaggle Result

```
#![Kaggle Result](https://github.com/angus001/Data605/blob/master/kagglefirsttry.PNG?raw=true)
```