Fundamentals of Computational Mathematics -605-Final Exam

Chirag Vithalani

December 10, 2016

Pick one of the quanititative independent variables from the training data set (train.csv) , and define that variable as X. Make sure this variable is skewed to the right! Pick the dependent variable and define it as Y.

Probability. Calculate as a minimum the below probabilities a through d. Assume the small letter "x" is estimated as the 3d quartile of the X variable, and the small letter "y" is estimated as the 2d quartile of the Y variable. Interpret the meaning of all probabilities. In addition, make a table of counts as shown below.

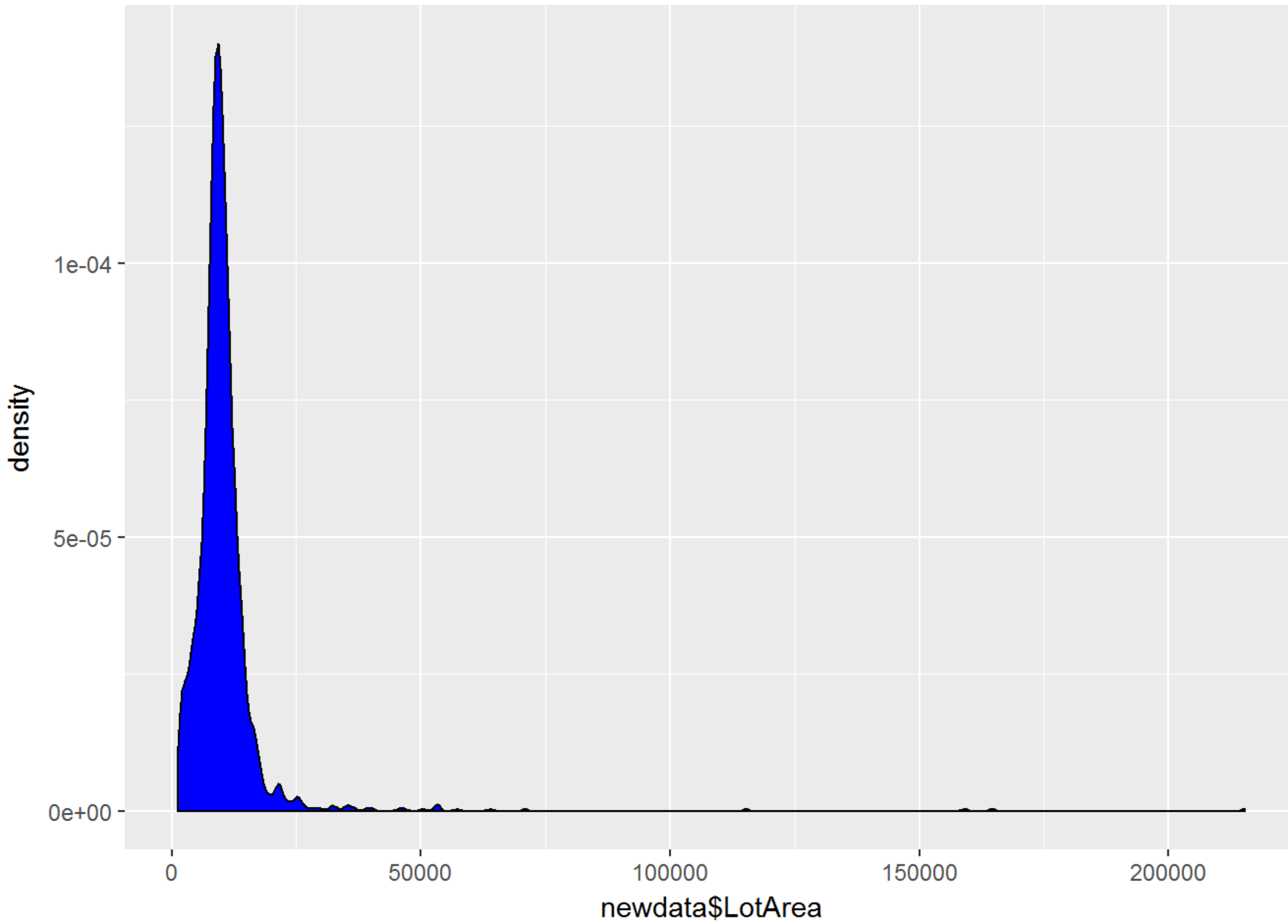
a. P(X>x | Y>y) b. P(X>x, Y>y) c. P(Xy)

Reading data from train.csv

Pick one of the quanititative independent variables from the training data set (train.csv) , and define that variable as X. Make sure this variable is skewed to the right! Pick the dependent variable and define it as Y.

I have chosen LotArea as independent variable and SalePrice as dependent variable.

```
#install.packages('ggplot2')
library(ggplot2)
exam_data<-read.csv("train.csv")
#economyRanking<-subset(exam_data, select=c("YearBuilt","SalePrice"))
economyRanking<-exam_data
#head(economyRanking,3)
newdata <- economyRanking[with(economyRanking, order(YearBuilt)), ]
#head(newdata)
ggplot(data = newdata) + geom_density(aes(x=newdata$LotArea), fill="blue")
```




```
#GarageArea
```

As we can see variable LotArea is right skewed (also known as positively skewed).

```
summary(newdata$LotArea)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1300    7554    9478   10520   11600  215200
```

Also for righ skewed data mean > media, which is confirmed.

a. P(X>x | Y>y)

Assume the small letter “x” is estimated as the 3d quartile of the X variable, and the small letter “y” is estimated as the 2d quartile of the Y variable.

```
x<-newdata$LotArea
y<-newdata$SalePrice
quantile(x, probs = 0.75,na.rm=TRUE)
```

```
##     75%
## 11601.5
```

```
quantile(y, probs = 0.5)
```

```
##    50%
## 163000
```

```
# Probability P(X > x and Y > y)
p1 <- nrow(subset(newdata, newdata$LotArea > quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE)
& newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE))) / nrow(newdata)


# Probability P(Y > y)
p2 <- nrow(subset(newdata, newdata$LotArea > quantile(newdata$LotArea, probs = 0.5,na.rm=TRUE)))
 / nrow(newdata)
#a.  P(X>x | Y>y)
p1 / p2
```

```
## [1] 0.3780822
```

b. P(X>x, Y>y)

```
nrow(subset(newdata, newdata$LotArea > quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE) & newd
ata$SalePrice > quantile(newdata$SalePrice, probs = 0.5))) / nrow(newdata)
```

```
## [1] 0.1890411
```

c. P(X < x | Y > y)

```
# Is the P(X < x and Y > y) divided by P(Y > y)

# Probability P(X < x and Y > y)
p1 <- nrow(subset(newdata, newdata$LotArea <= quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE)
 & newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE))) / nrow(newdata)

# Probability P(Y > y)
p2 <- nrow(subset(newdata, newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.5,na.rm=TRU
E))) / nrow(newdata)
p1 / p2
```

```
## [1] 0.6208791
```

```
# Compute value for (a)
a<-nrow(subset(newdata, newdata$LotArea <= quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE) &
newdata$SalePrice <= quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE)))
a
```

```
## [1] 643
```

```
# Compute value for (b)
b<-nrow(subset(newdata, newdata$LotArea <= quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE) &
newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE)))
b
```

```
## [1] 452
```

```
# Compute value for (c)
c<-nrow(subset(newdata, newdata$LotArea > quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE) & n
ewdata$SalePrice <= quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE)))

c
```

```
## [1] 89
```

```
# Compute value for (d)
d<-nrow(subset(newdata, newdata$LotArea > quantile(newdata$LotArea, probs = 0.75,na.rm=TRUE) & n
ewdata$SalePrice > quantile(newdata$SalePrice, probs = 0.5,na.rm=TRUE)))

d
```

```
## [1] 276
```

x/y	<=2d quartile	>2d quartile	Total
<=3d quartile	643	452	1095
>3d quartile	89	276	365
Total	732	728	1460

Does splitting the data in this fashion make them independent? No. The fact that we can take observations and subset them doesn't make them independent or dependent.

Let A be the new variable counting those observations above the 3rd quartile for X, and let B be the new variable counting those observations for the 2nd quartile for Y. Does P(A|B) = P(A) * P(B)? Check mathematically. No - see below.

```
A <- nrow(subset(newdata, newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.75,na.rm=TRU
E)))
B <- nrow(subset(newdata, newdata$LotArea <= quantile(newdata$LotArea, probs = 0.5,na.rm=TRUE)))
# P(A)
pA <- A / nrow(newdata)
# P(B)
pB <- B / nrow(newdata)
# P(A|B)
pAB <- nrow(subset(newdata, newdata$SalePrice > quantile(newdata$SalePrice, probs = 0.75,na.rm=T
RUE) & newdata$LotArea <= quantile(newdata$LotArea, probs = 0.5,na.rm=TRUE))) / nrow(newdata)

pA * pB
```

```
## [1] 0.1239726
```

```
pAB
```

```
## [1] 0.04794521
```

Evaluate by running a Chi Square test for association.

```
chisqtbl <- table(newdata$SalePrice, newdata$LotArea)
chisq.test(chisqtbl)
```

```
## Warning in chisq.test(chisqtbl): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisqtbl
## X-squared = 735090, df = 709660, p-value < 2.2e-16
```

Descriptive and Inferential Statistics

Provide univariate descriptive statistics and appropriate plots for the training data set.Provide a scatterplot of X and Y. Provide a 95% CI for the difference in the mean of the variables.Derive a correlation matrix for two of the quantitative variables you selected.Test the hypothesis that the correlation between these variables is 0 and provide a 99% confidence interval.Discuss the meaning of your analysis.

Here are summary statistics for SalePrice and LotArea to supplment the Histograms in the prior section:

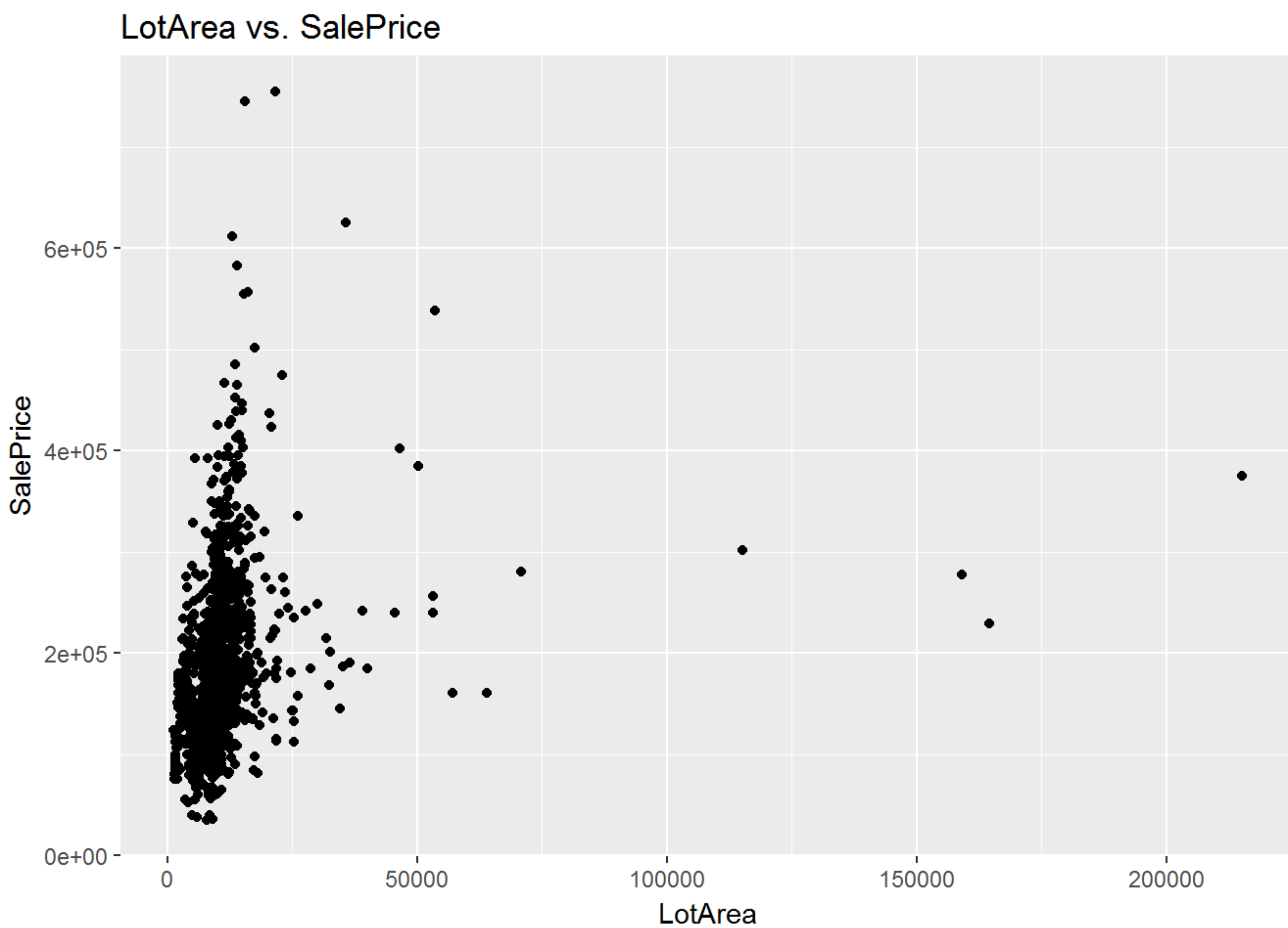
```
summary(newdata$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  130000  163000  180900  214000  755000
```

```
summary(newdata$LotArea)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1300    7554    9478   10520   11600  215200
```

```
# Load ggplot2
#install.packages("ggplot2", repos='https://mirrors.nics.utk.edu/cran/')
library(ggplot2)

ggplot(newdata, aes(x=newdata$LotArea, y=newdata$SalePrice)) + geom_point() + labs(title="LotAre
a vs. SalePrice",x="LotArea", y = "SalePrice")
```



scatterplot

```
#ggplot(newdata, aes(x=newdata$LotArea, y=newdata$SalePrice)) + geom_point() + labs(title="LotAr
ea vs. SalePrice",x="LotArea", y = "SalePrice")
```

It certainly looks like the variables are correlated.

Provide a 95% confidence interval for the difference in the mean of the variables.

```
# Difference between the means
dm <- mean(newdata$SalePrice) - mean(newdata$LotArea,na.rm = TRUE)
dm
```

```
## [1] 170404.4
```

```
# Standard error of the difference between means
se <- sqrt(((sd(newdata$SalePrice)/nrow(newdata))+(sd(newdata$LotArea,na.rm =
TRUE)/nrow(newdata))))
se
```

```
## [1] 7.826184
```

```
# 95% confidence interval
c(dm - se*qnorm(0.975),dm + se*qnorm(0.975))
```

```
## [1] 170389.0 170419.7
```

Derive a correlation matrix for two of the quantitative variables you selected.

```
newdata1<-subset(newdata, LotArea !='NA')
newdata<-newdata1

cm <- cor(newdata[c("SalePrice","LotArea")])
cm
```

```
##           SalePrice   LotArea
## SalePrice 1.0000000 0.2638434
## LotArea   0.2638434 1.0000000
```

Test the hypothesis that the correlation between these variables is 0 and provide a 99% confidence interval.

```
cor.test(newdata$SalePrice,newdata$LotArea,conf.level = 0.99)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  newdata$SalePrice and newdata$LotArea
## t = 10.445, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  0.2000196 0.3254375
## sample estimates:
##       cor
## 0.2638434
```

Given the p-value, the likelihood that the hypothesis of a zero correlation is very low.

Linear Algebra and Correlation

Invert your correlation matrix. (This is known as the precison matrix).

```
im <- solve(cm)
im
```

```
##             SalePrice    LotArea
## SalePrice  1.0748219 -0.2835846
## LotArea   -0.2835846  1.0748219
```

** Multiply the correlation matrix by the precision matrix, and then mutiply the precision matrix by the correlation matrix. **

```
cm %*% im
```

```
##           SalePrice LotArea
## SalePrice         1       0
## LotArea           0       1
```

```
im %*% cm
```

```
##           SalePrice LotArea
## SalePrice         1       0
## LotArea           0       1
```

The result in both cases is the identity matrix.

Calculus Based Probability and Statistics

For your variable which is skewed to the right, shift it so that the minimum value is above zero.

```
#Create new DF
hf_min_val <- newdata
# Check range for SalePrice
c(hf_min_val$SalePrice[which.min(hf_min_val$SalePrice)],hf_min_val$SalePrice[which.max(hf_min_va
l$SalePrice)])
```

```
## [1]  34900 755000
```

```
# Add 34 to all values
#hf_min_val$SalePrice <- hf_min_val$SalePrice + 34
# Check range for LotArea
c(hf_min_val$LotArea[which.min(hf_min_val$LotArea)],hf_min_val$LotArea[which.max(hf_min_val$LotA
rea)])
```

```
## [1]   1300 215245
```

```
# Add 71 to all values
#hf_min_val$LotArea <- hf_min_val$LotArea + 71
```

Though both variabiles are skewed to the right, for this exercise, I will use the LotArea variable.

Then load the MASS package and run fitdistr to fit an exponential probability density function. Documentation for MASS::fitdistr is here: https::/stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html (https::/stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html)

```
install.packages("MASS", repos='https://mirrors.nics.utk.edu/cran/')
```

```
## package 'MASS' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chirag.vithalani\AppData\Local\Temp\RtmpATJZr1\downloaded_packages
```

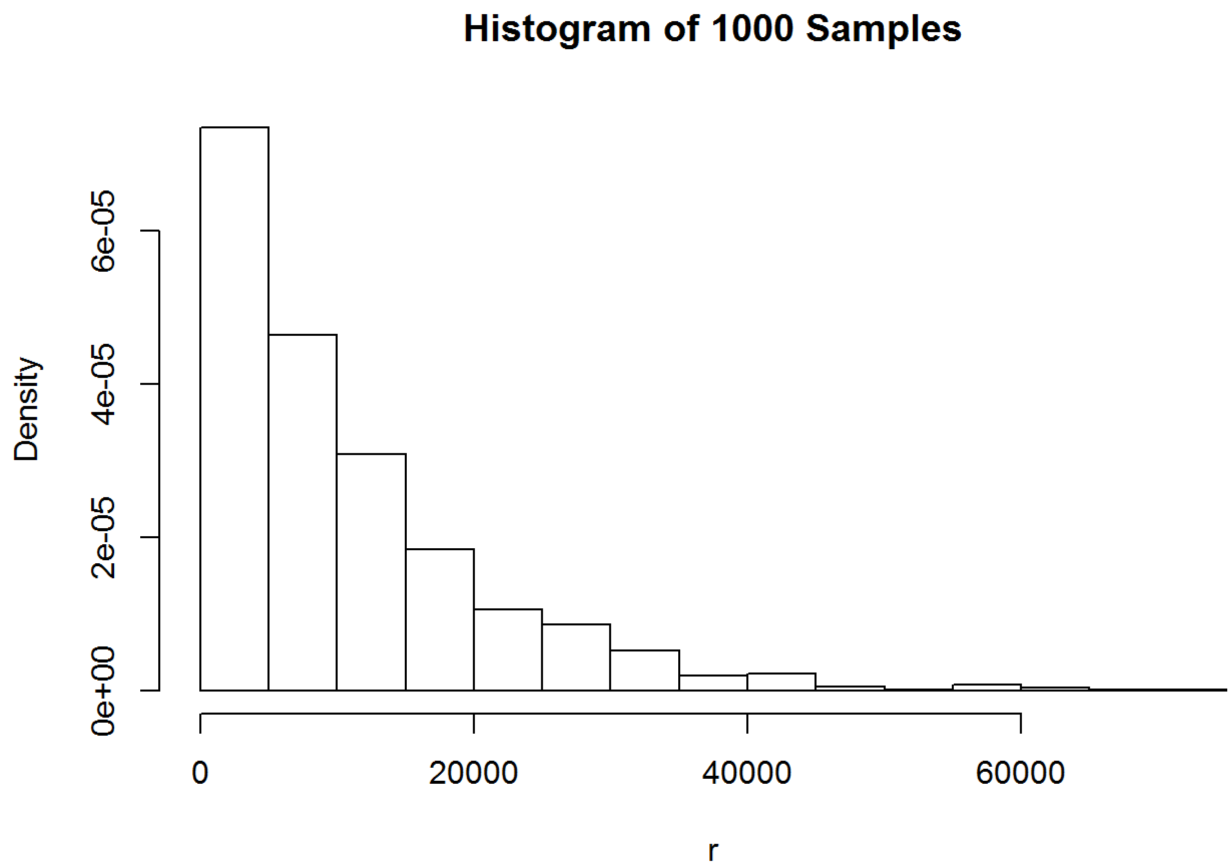
```
library(MASS)
fd <- fitdistr(hf_min_val$LotArea, "exponential")
```

Find the optimal value of lambda for this distribution, and then take 1000 samples from this exponential distribution using this value. Plot a histogram and compare it with a histogram of your original variable.

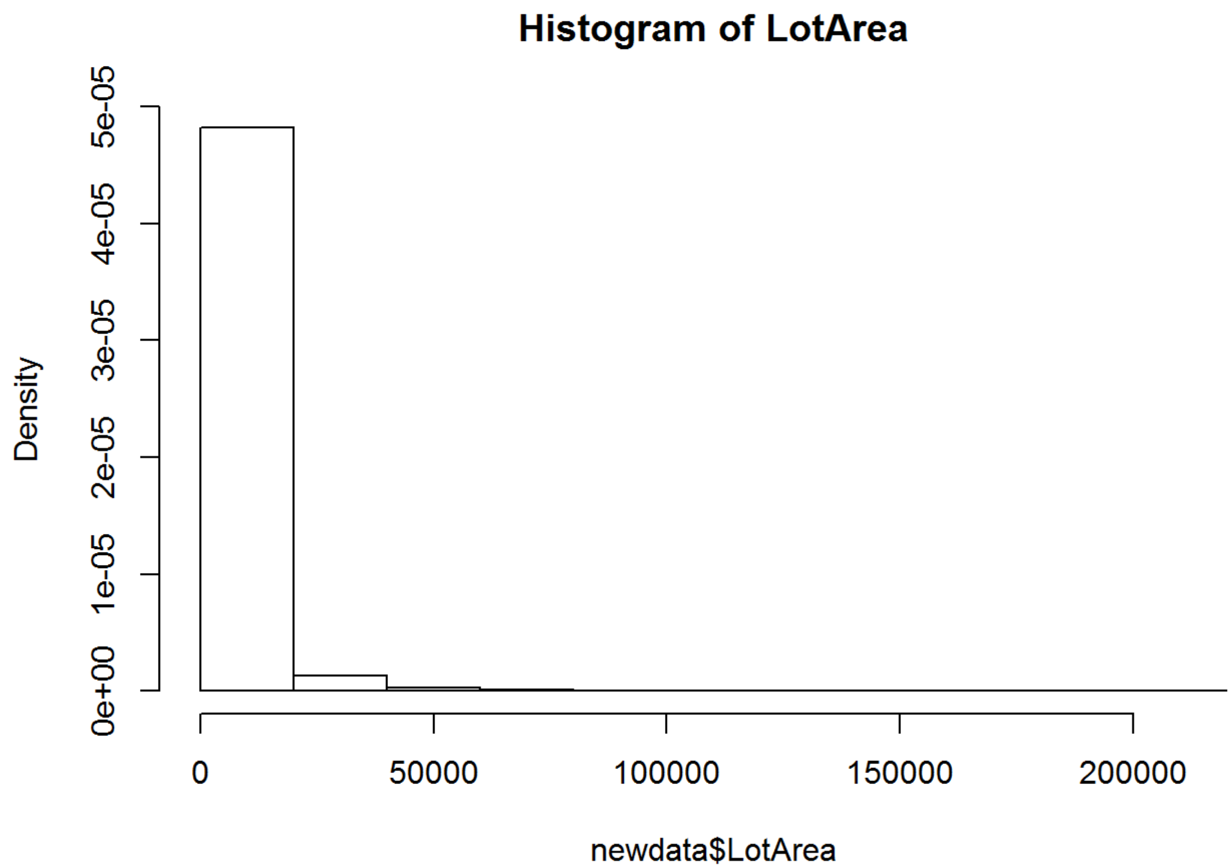
```
# Optimal value of lambda
fd$estimate
```

```
##        rate
## 9.50857e-05
```

```
# 100 samples from this distribution
r <- rexp(1000,fd$estimate)
# Plot a histogram using 1000 samples
hist(r, freq = FALSE, main = "Histogram of 1000 Samples")
```




```
# Plot a histogram using original variable
hist(newdata$LotArea, freq = FALSE,  main = "Histogram of LotArea")
```



Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).

```
# 5th percentile
pexp(quantile(r, probs = 0.05), rate=fd$estimate, lower.tail = TRUE)
```

```
##         5%
## 0.05276064
```

```
# 95th percentile
pexp(quantile(r, probs = 0.95), rate=fd$estimate, lower.tail = TRUE)
```

```
##       95%
## 0.9508811
```

Also generate a 95% confidence interval from the empirical data, assuming normality.

```
# y was set to newdata$LotArea earlier
# Calculate mean and standard deviation
m <- mean(y)
se <- sd(y)
# 95% confidence interval
c(m - (se*qnorm(0.975)), m + (se*qnorm(0.975)))
```

```
## [1]  25216.75 336625.64
```

Finally, provide the empirical 5th percentile and 95th percentile of the data.

```
quantile(y, probs = 0.05)
```

```
##    5%
## 88000
```

```
quantile(y, probs = 0.95)
```

```
##    95%
## 326100
```

Discuss.

There is a large divergence between the 5% and 95% percentiles (assuming normality) and the 5th and 95th percentiles determined empirically because the distribution is not normal, it is exponetial.

Modeling.

Build some type of regression model and submit your model to the competition board.

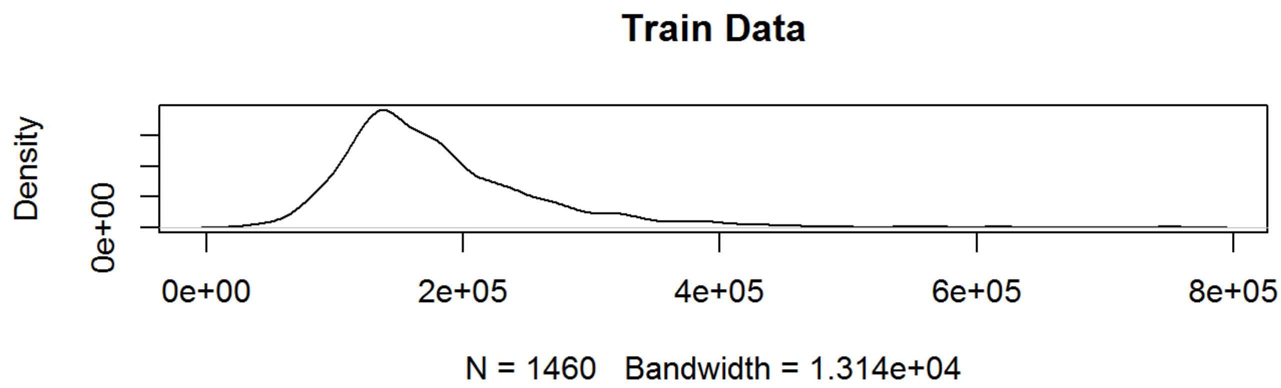
```
#reading test.csv file
test <-read.csv("https://raw.githubusercontent.com/chirag-vithlani/Fundamentals-of-Computational
-Mathematics-605/master/project/test.csv", header = TRUE)

#reading train.csv file
train <- read.csv("https://raw.githubusercontent.com/chirag-vithlani/Fundamentals-of-Computation
al-Mathematics-605/master/project/train.csv", header = TRUE)

train<-subset(train, select=c("Id", "LotArea","SalePrice"))

#Removing ID column
train<-train%>%select(-Id)


traincleaned<-cleanup(train)
testcleaned<-cleanup(test)

# Fitting Linear Models
upper<-lm(SalePrice~.,traincleaned)
lower<-lm(SalePrice~1, traincleaned)

stepResults<-step(lower,scope=list(lower=lower,upper=upper),direction="both")
```

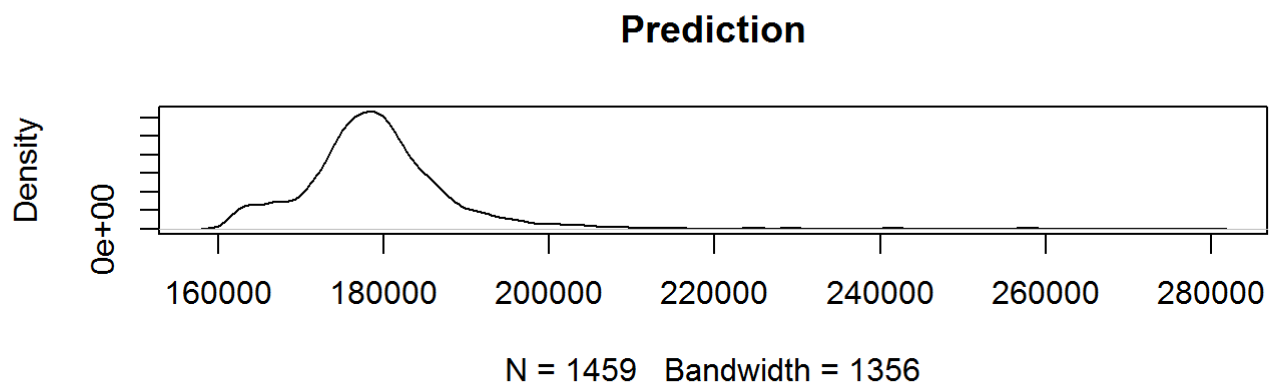
```
## Start:  AIC=32946.74
## SalePrice ~ 1
##
##           Df  Sum of Sq        RSS   AIC
## + LotArea  1 6.4099e+11 8.5669e+12 32843
## <none>                  9.2079e+12 32947
##
## Step:  AIC=32843.4
## SalePrice ~ LotArea
##
##           Df  Sum of Sq        RSS   AIC
## <none>                  8.5669e+12 32843
## - LotArea  1 6.4099e+11 9.2079e+12 32947
```

```
par(mfrow=c(1,1))
#plot(data.frame('SalePrice'=predict(stepResults, testcleaned), 'Id'=testcleaned$Id))

result<-data.frame('Id'=testcleaned$Id,'SalePrice'=predict(stepResults, testcleaned))
length(result$SalePrice)
```

```
## [1] 1459
```

```
write.csv(result, file = "kaggle_Chirag.csv", row.names = F)

par(mfrow=c(2,1))
plot(density(traincleaned$SalePrice),main="Train Data")
plot(density(na.omit(result$SalePrice)),main="Prediction")
```





Below shows the score on Kaggle