# Author: Brian Pankau

# Class: UWLAX DS700

# Assignment: 11 R

# ----------------------------------------------------------------------------

# PROBLEM 1E

# Two records (44727 64412) were found in the data frame where the total number

# votes were less than the votes voted as helpful which is an unrealistic

# scenario for the ratio of votes helpful / votes total > 1.

# ----------------------------------------------------------------------------

# PROBLEM 1F

OBSERVATION: There were 270054 missing values for helpful reviews, of which 270052 values were due to no review feedback ("0/0"); the remaining 2 observations were the two unrealistic records.

# ----------------------------------------------------------------------------

# PROBLEM 1G: Review length summary statistics

**ANALYSIS:**

Question: Are helpful reviews longer than unhelpful ones?

Answer: As depicted by the summary descriptive statistics, the helpful reviews have more review texts that are of longer length than the reiews that are not helpful (see, for example, the quantile reports). However, as illustrated by the boxplot distributions, there are a lot of outliers; the presence of which is transforming the distribution curve to a non-normal distribution. An examination of the grouped frequency distributions indicates that the length of the review texts for both the helpful reviews and the unhelpful reviews are skewed to the right; also presenting a non-normal distribution, indicating a need to apply a transformation on the data, such as a square root, log, or inverse log transformation (all transformations ploted, see below), and/or binning (see the grouping plots); or there may be a need to resort to applying non-parametric techniques against the data.

A comparison of the raw and transformed plots of the revew length indicate that an log transformation significantly normalizes the distribution. The hypothesis tests will use a log transformation to compare the means between the lengths of the helpful reviews and the unhelpful reviews.

Test the hypothesis that the lengths of the helpful reviews are longer to the length of the unhelpful reviews. Use a 0.05 level of significance.

**Testing Assumptions:**

* Apply an inverse log transformation on the length of the review text to attempt to conform the distribution into an approximate normal distribution.
* Samples for the lengths of the reviews are from a random sample.
* The helpful reviews are independent from the unhelpful reviews; indicating a non-paired type of test.
* Since we are trying to determine if helpful reviews are longer than unhelpful reviews, this is a greater than or equal to test case, so apply a two-tailed test.
* Since random sample of a 1000 observations of the two review types have equal variances (as confirmed by the variance F-test), a Welch Two Sample t-test will be applied rather than the Wilcoxon-Mann-Whitney test.
* Rejection criteria: If the p-value returned by the t-test is less than 0.05, then reject the null hypothesis in favor of the alternate hypothesis.

Null Hypothesis (H0): Using an alpha of 0.05 level of significance, the mean of the lengths of helpful reviews are the same as unhelpful reviews. (µHelpful ≤ µUnhelpful)

Alternate Hypothesis (H1): Using an alpha of 0.05 level of significance, the mean of the lengths of helpful reviews are longer than unhelpful reviews.(µHelpful › µUnhelpful)

**Perform Test**

# generate a random sample of 1000 for each review type

# since there is an unequal number of total reviews between the two review types

randomSample = function(df,n) {

return (df[sample(nrow(df), n),])

}

df\_helpful\_reviews\_sample <- randomSample(df\_helpful\_reviews, 1000)

df\_unhelpful\_reviews\_sample <- randomSample(df\_unhelpful\_reviews, 1000)

setup variables

x <- df\_helpful\_reviews\_sample$Review\_length\_log

y <- df\_unhelpful\_reviews\_sample$Review\_length\_log

alpha\_value <- 0.05

Evaluate same variances of the two groups to verify homoskedasticity

> var.test(a,b)

F test to compare two variances

data: x and y

F = 0.95206, num df = 999, denom df = 999, p-value = 0.4376

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.8409505 1.0778424

sample estimates:

ratio of variances

0.9520568

Findings for the means variance F-test: The F-Test p-value is > 0.05 indicating that the variances between the type review types are homogeneous indicating that it is acceptable to apply the Welch Two Sample t-test rather than the Wilcoxon-Mann-Whitney test.

Perform Student T-Test:

> t.test(x, y, alternative = "greater")

Welch Two Sample t-test

data: x and y

t = 5.324, df = 1996.8, p-value = 5.646e-08

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.1262225 Inf

sample estimates:

mean of x mean of y

5.907599 5.724908

Findings for the Welch Two Sample t-test:

The p-value returned by the Welch Two Sample t-test is less than 0.05 (and which is less than the level of significance used in this scenario’s null hypothesis), and so the null hypothesis is rejected in favor of the alternate hypothesis that the means of the lengths of the helpful reveiws is longer than the means of the lengths of the unhelpful review.

**Descriptive Statistics for Helpful Reviews**

> summary(df\_helpful\_reviews$Review\_length)

Min. 1st Qu. Median Mean 3rd Qu. Max.

30 202 347 500 604 21221

> Mode(df\_helpful\_reviews$Review\_length)

[1] "165"

> var(df\_helpful\_reviews$Review\_length)

[1] 269000

> sd(df\_helpful\_reviews$Review\_length)

[1] 519

> range(df\_helpful\_reviews$Review\_length)

[1] 30 21221

> quantile(df\_helpful\_reviews$Review\_length)

0% 25% 50% 75% 100%

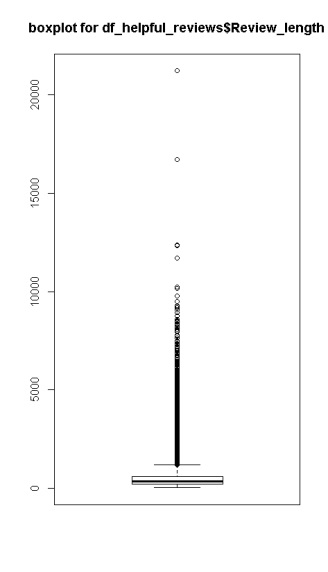
30 202 347 604 21221

> mad(df\_helpful\_reviews$Review\_length)

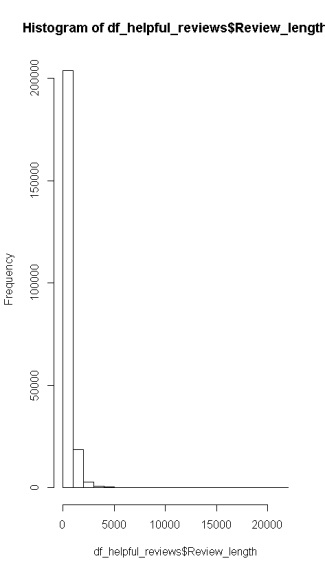
[1] 256

**Non-Transformed Non-Binned Initial Plots for Helpful Reviews**

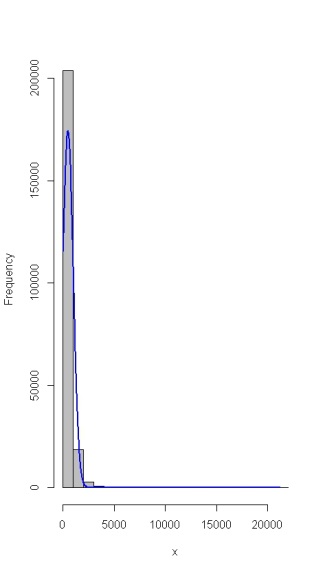
boxplot(df\_helpful\_reviews$Review\_length, main="boxplot for df\_helpful\_reviews$Review\_length")



hist(df\_helpful\_reviews$Review\_length)



plotNormalHistogram(df\_helpful\_reviews$Review\_length)



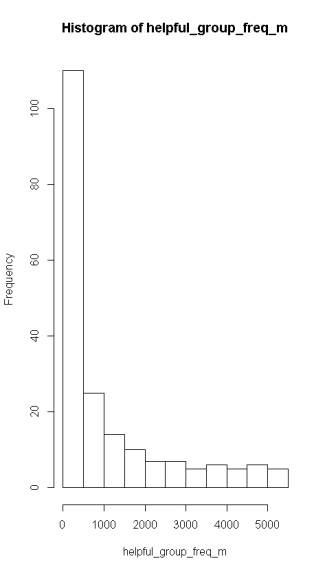
**Non-Transformed Binned Initial Plots for Helpful Reviews**

helpful\_breaks <- seq(0,2000, by=10)

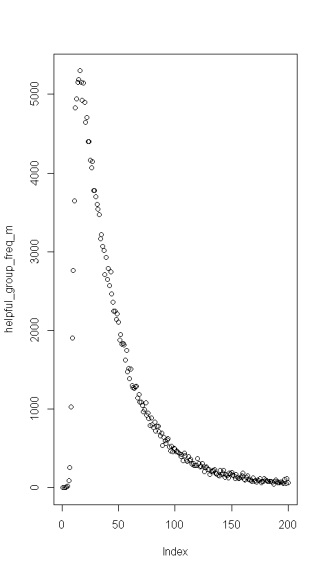
helpful\_group\_freq <- cut(df\_helpful\_reviews$Review\_length, helpful\_breaks, right=FALSE)

helpful\_group\_freq\_m <- cbind(table(helpful\_group\_freq))

hist(helpful\_group\_freq\_m)

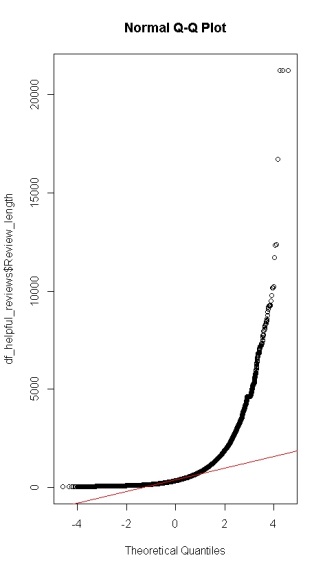


plot(helpful\_group\_freq\_m)



qqnorm(df\_helpful\_reviews$Review\_length, ylab="df\_helpful\_reviews$Review\_length")

qqline(df\_helpful\_reviews$Review\_length, col="red")



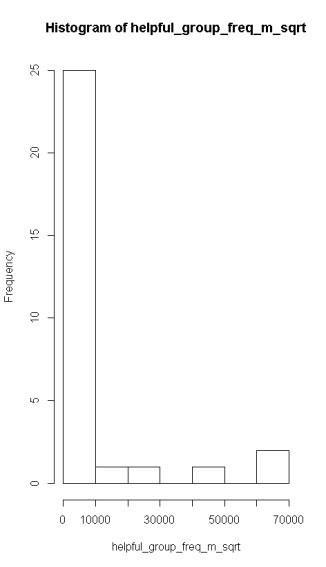
**Transformed Binned Distribution Plots for Helpful Reviews**

range(df\_helpful\_reviews$Review\_length\_sqrt); helpful\_breaks\_sqrt <- seq(0,150, by=5)

helpful\_group\_freq\_sqrt <- cut(df\_helpful\_reviews$Review\_length\_sqrt, helpful\_breaks\_sqrt, right=FALSE)

helpful\_group\_freq\_m\_sqrt <- cbind(table(helpful\_group\_freq\_sqrt))

hist(helpful\_group\_freq\_m\_sqrt)

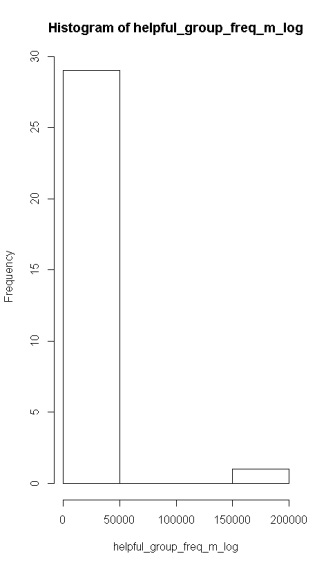


range(df\_helpful\_reviews$Review\_length\_log); helpful\_breaks\_log <- seq(0,150, by=5)

helpful\_group\_freq\_log <- cut(df\_helpful\_reviews$Review\_length\_log, helpful\_breaks\_log, right=FALSE)

helpful\_group\_freq\_m\_log <- cbind(table(helpful\_group\_freq\_log))

hist(helpful\_group\_freq\_m\_log)

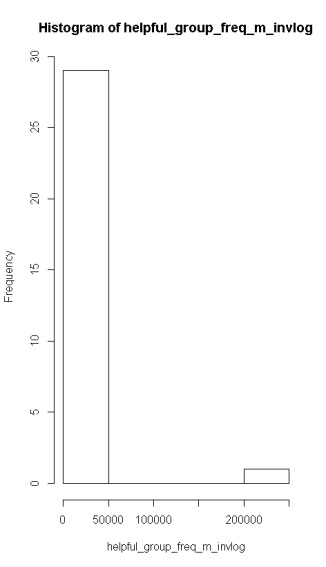


range(df\_helpful\_reviews$Review\_length\_invlog); helpful\_breaks\_invlog <- seq(0,150, by=5)

helpful\_group\_freq\_invlog <- cut(df\_helpful\_reviews$Review\_length\_invlog, helpful\_breaks\_invlog, right=FALSE)

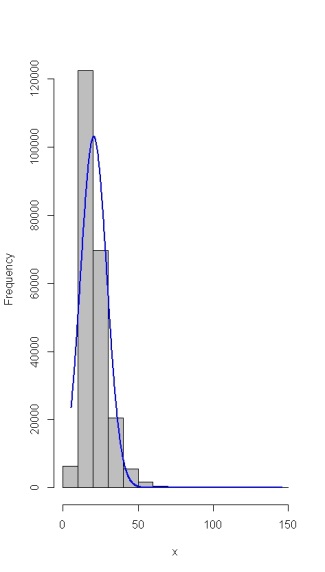
helpful\_group\_freq\_m\_invlog <- cbind(table(helpful\_group\_freq\_invlog))

hist(helpful\_group\_freq\_m\_invlog)

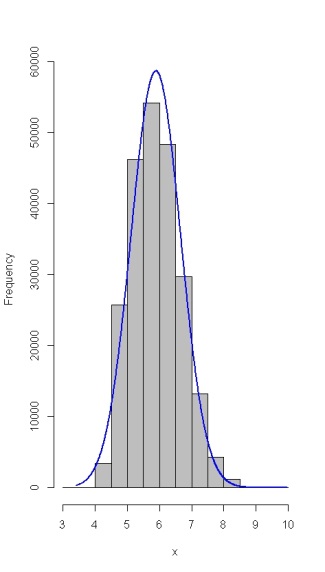


**Transformed Non-Binned Histogram + Normal Distribution for Helpful Reviews**

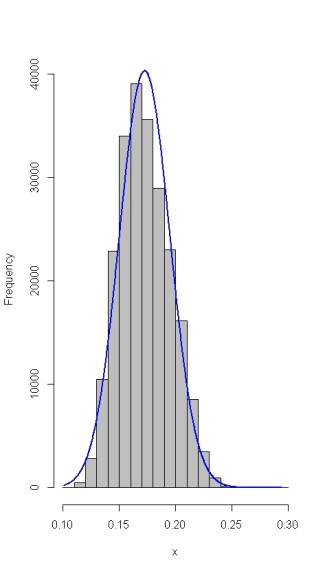
plotNormalHistogram(df\_helpful\_reviews$Review\_length\_sqrt)



plotNormalHistogram(df\_helpful\_reviews$Review\_length\_log)



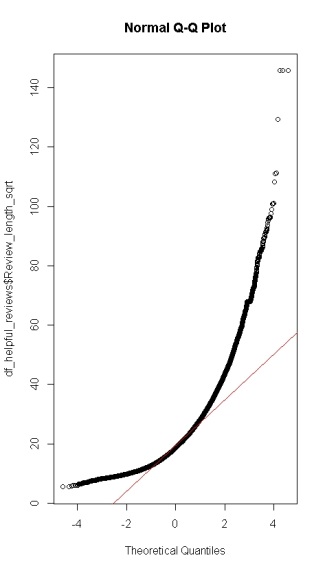
plotNormalHistogram(df\_helpful\_reviews$Review\_length\_invlog)



**Transformed Non-Binned Q-Q Plots for Helpful Reviews**

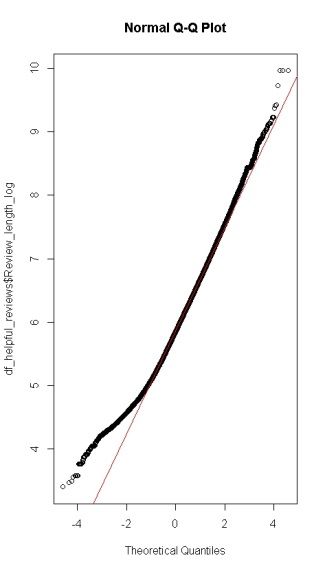
qqnorm(df\_helpful\_reviews$Review\_length\_sqrt, ylab="df\_helpful\_reviews$Review\_length\_sqrt")

qqline(df\_helpful\_reviews$Review\_length\_sqrt, col="red")



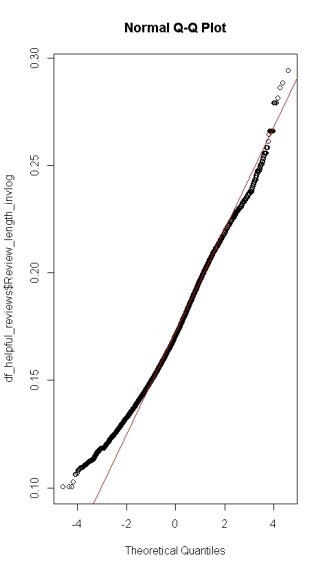
qqnorm(df\_helpful\_reviews$Review\_length\_log, ylab="df\_helpful\_reviews$Review\_length\_log")

qqline(df\_helpful\_reviews$Review\_length\_log, col="red")



qqnorm(df\_helpful\_reviews$Review\_length\_invlog, ylab="df\_helpful\_reviews$Review\_length\_invlog")

qqline(df\_helpful\_reviews$Review\_length\_invlog, col="red")



**Descriptive Statistics for Unhelpful Reviews**

> summary(df\_unhelpful\_reviews$Review\_length)

Min. 1st Qu. Median Mean 3rd Qu. Max.

12 171 298 442 527 13047

> Mode(df\_unhelpful\_reviews$Review\_length)

[1] "124"

> var(df\_unhelpful\_reviews$Review\_length)

[1] 236116

> sd(df\_unhelpful\_reviews$Review\_length)

[1] 486

> range(df\_unhelpful\_reviews$Review\_length)

[1] 12 13047

> quantile(df\_unhelpful\_reviews$Review\_length)

0% 25% 50% 75% 100%

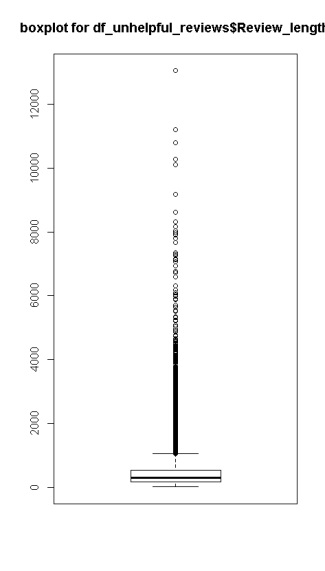
12 171 298 527 13047

> mad(df\_unhelpful\_reviews$Review\_length)

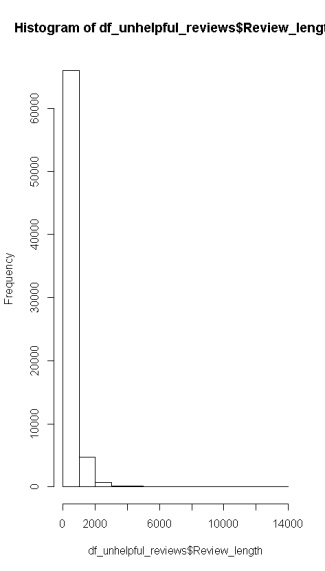
[1] 224

**Non-Transformed Non-Binned Initial Plots for Unhelpful Reviews**

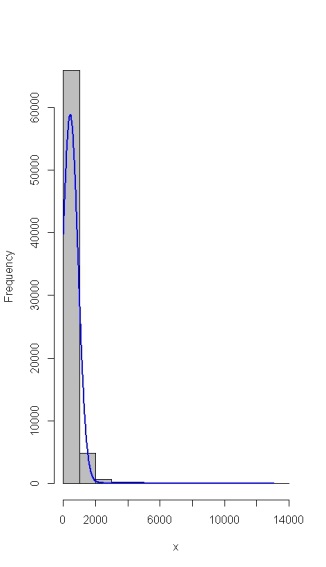
boxplot(df\_unhelpful\_reviews$Review\_length, main="boxplot for df\_unhelpful\_reviews$Review\_length")



hist(df\_unhelpful\_reviews$Review\_length)



plotNormalHistogram(df\_unhelpful\_reviews$Review\_length)



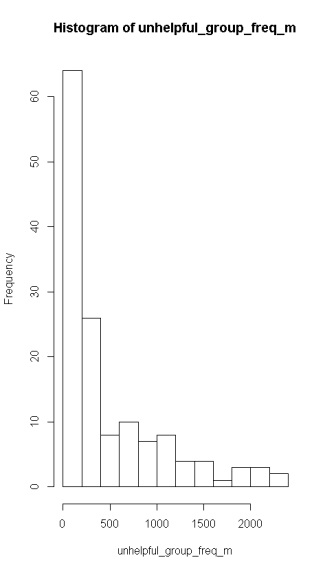
**Non-Transformed Binned Initial Plots for Unhelpful Reviews**

unhelpful\_breaks <- seq(0, 1400, by=10)

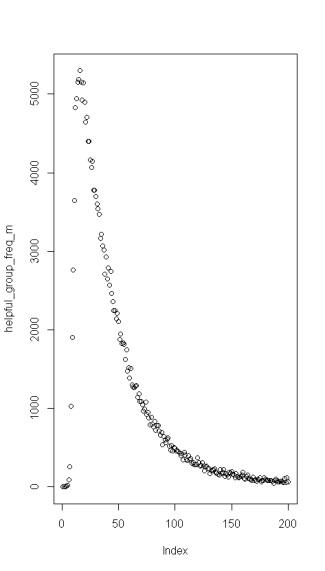
unhelpful\_group\_freq <- cut(df\_unhelpful\_reviews$Review\_length, unhelpful\_breaks, right=FALSE)

unhelpful\_group\_freq\_m <- cbind(table(unhelpful\_group\_freq))

hist(unhelpful\_group\_freq\_m)

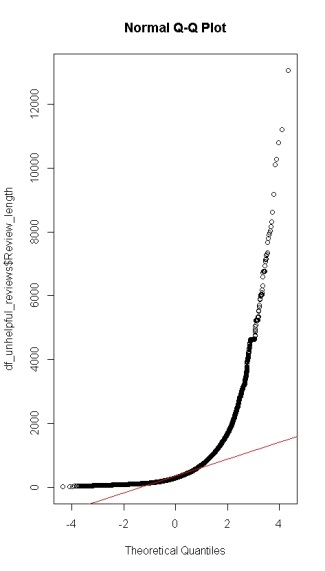


plot(unhelpful\_group\_freq\_m)



qqnorm(df\_unhelpful\_reviews$Review\_length, ylab="df\_unhelpful\_reviews$Review\_length")

qqline(df\_unhelpful\_reviews$Review\_length, col="red")



**Transformed Binned Distribution Plots for Helpful Reviews**

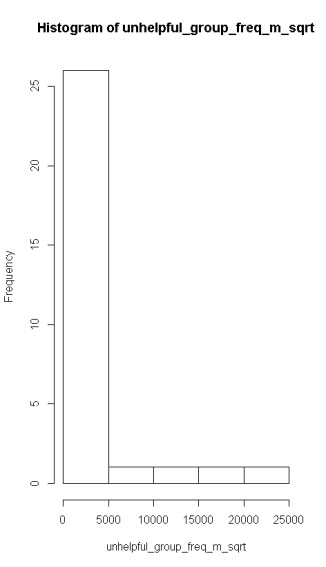
**Population**

range(df\_unhelpful\_reviews$Review\_length\_sqrt)

unhelpful\_breaks\_sqrt <- seq(0, 150, by=5)

unhelpful\_group\_freq\_sqrt <- cut(df\_unhelpful\_reviews$Review\_length\_sqrt, unhelpful\_breaks\_sqrt, right=FALSE)

unhelpful\_group\_freq\_m\_sqrt <- cbind(table(unhelpful\_group\_freq\_sqrt)); hist(unhelpful\_group\_freq\_m\_sqrt)

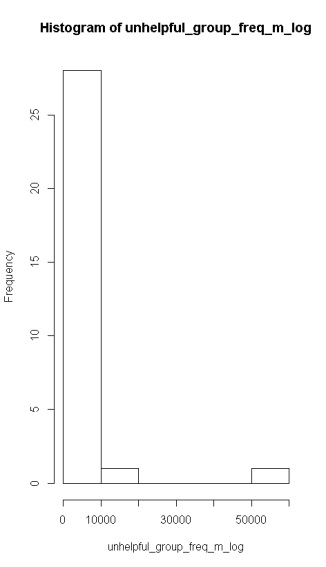


range(df\_unhelpful\_reviews$Review\_length\_log)

unhelpful\_breaks\_log <- seq(0,150, by=5)

unhelpful\_group\_freq\_log <- cut(df\_unhelpful\_reviews$Review\_length\_log, unhelpful\_breaks\_log, right=FALSE)

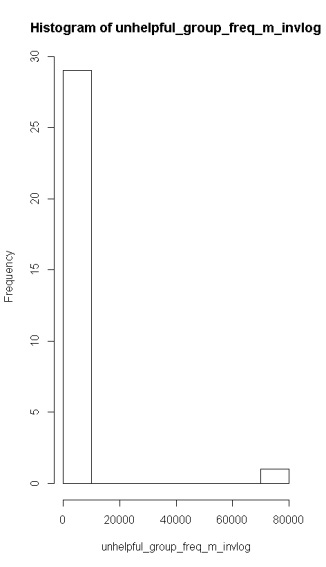
unhelpful\_group\_freq\_m\_log <- cbind(table(unhelpful\_group\_freq\_log)); hist(unhelpful\_group\_freq\_m\_log)



range(df\_unhelpful\_reviews$Review\_length\_invlog), unhelpful\_breaks\_invlog <- seq(0,150, by=5)

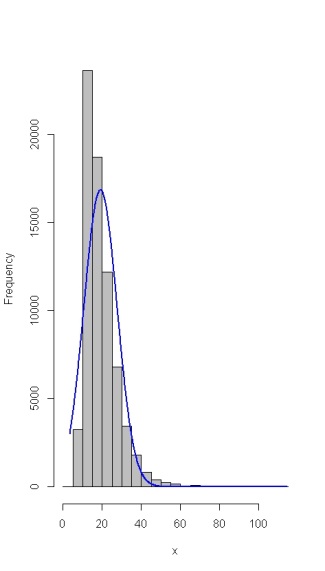
unhelpful\_group\_freq\_invlog <- cut(df\_unhelpful\_reviews$Review\_length\_invlog, unhelpful\_breaks\_invlog, right=FALSE)

unhelpful\_group\_freq\_m\_invlog <- cbind(table(unhelpful\_group\_freq\_invlog)); hist(unhelpful\_group\_freq\_m\_invlog)

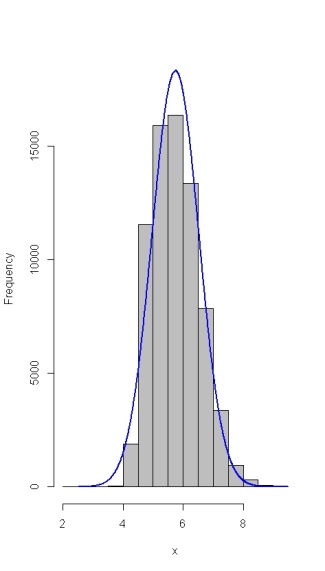


**Transformed Non-Binned Histogram + Normal Distribution for Unhelpful Reviews**

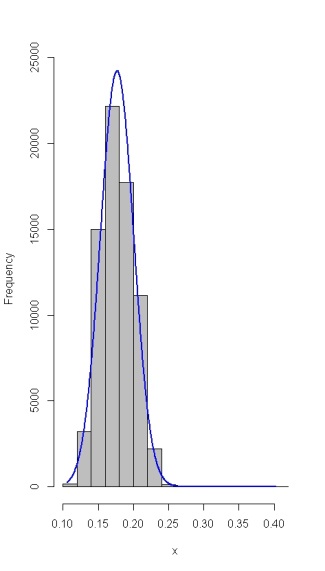
plotNormalHistogram(df\_unhelpful\_reviews$Review\_length\_sqrt)



plotNormalHistogram(df\_unhelpful\_reviews$Review\_length\_log)

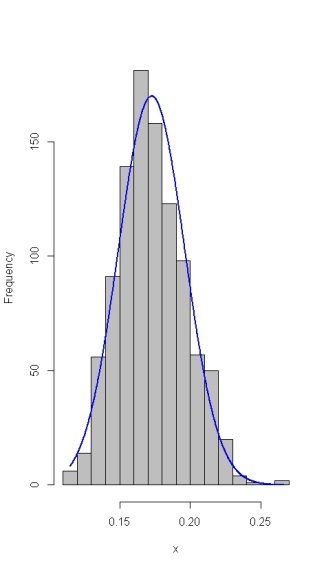


plotNormalHistogram(df\_unhelpful\_reviews$Review\_length\_invlog)



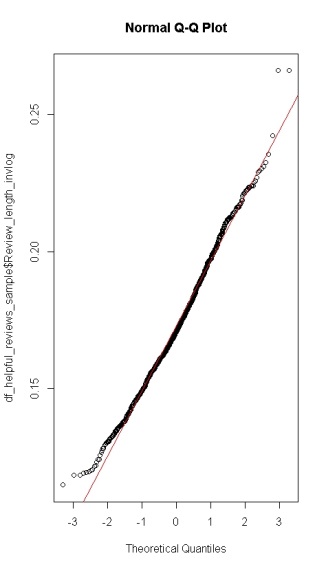
**Sample**

plotNormalHistogram(df\_helpful\_reviews\_sample$Review\_length\_invlog)



qqnorm(df\_helpful\_reviews\_sample$Review\_length\_invlog, ylab="df\_helpful\_reviews\_sample$Review\_length\_invlog")

qqline(df\_helpful\_reviews\_sample$Review\_length\_invlog, col="red")

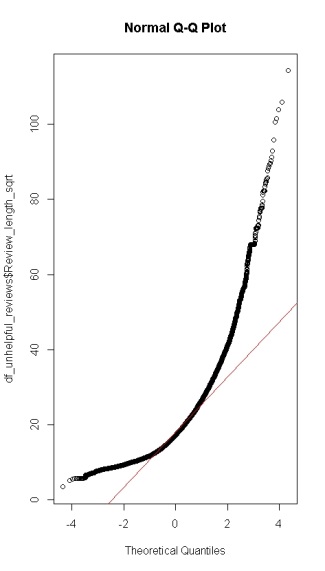


**Transformed Non-Binned Q-Q Plots for Unhelpful Reviews**

**Population**

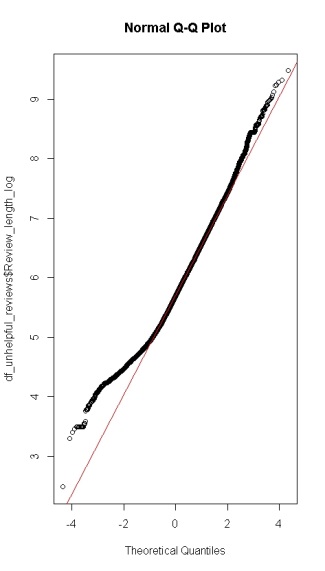
qqnorm(df\_unhelpful\_reviews$Review\_length\_sqrt, ylab="df\_unhelpful\_reviews$Review\_length\_sqrt")

qqline(df\_unhelpful\_reviews$Review\_length\_sqrt, col="red")



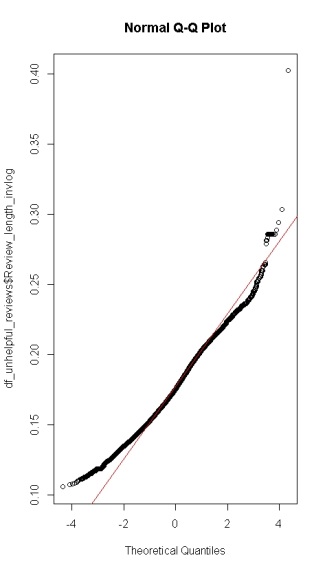
qqnorm(df\_unhelpful\_reviews$Review\_length\_log, ylab="df\_unhelpful\_reviews$Review\_length\_log")

qqline(df\_unhelpful\_reviews$Review\_length\_log, col="red")



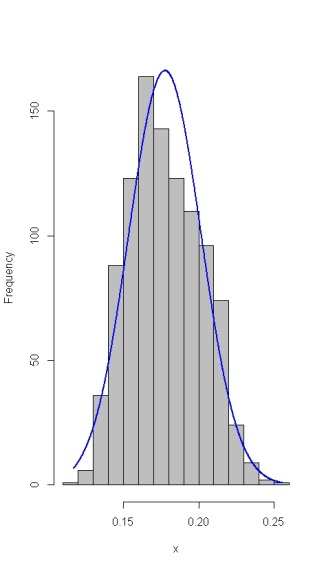
qqnorm(df\_unhelpful\_reviews$Review\_length\_invlog, ylab="df\_unhelpful\_reviews$Review\_length\_invlog")

qqline(df\_unhelpful\_reviews$Review\_length\_invlog, col="red")



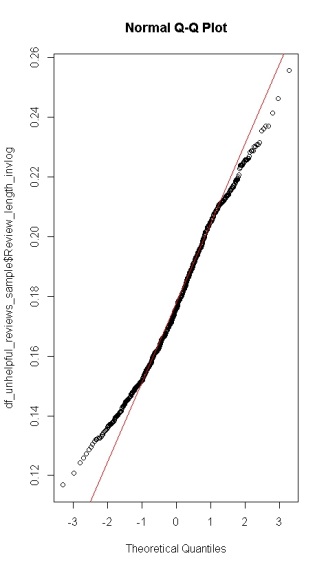
**Sample**

plotNormalHistogram(df\_unhelpful\_reviews\_sample$Review\_length\_invlog)



qqnorm(df\_unhelpful\_reviews\_sample$Review\_length\_invlog, ylab="df\_unhelpful\_reviews\_sample$Review\_length\_invlog")

qqline(df\_unhelpful\_reviews\_sample$Review\_length\_invlog, col="red")



# ----------------------------------------------------------------------------

# PROBLEM 1H

# find all unique ProductID

ProductID\_unique <- unique(df$ProductId)

> head(ProductID\_unique)

[1] B001E4KFG0 B00813GRG4 B000LQOCH0 B000UA0QIQ B006K2ZZ7K B000E7L2R4

74258 Levels: 0006641040 141278509X 2734888454 2841233731 7310172001 7310172101 7800648702 9376674501 B00002N8SM B00002NCJC ... B009WVB40S

# count the number of reviews per unique ProductID

ProductID\_votedtotal\_count <- tapply(df$Voted\_total, df$ProductId, length)

head(ProductID\_votedtotal\_count)

0006641040 141278509X 2734888454 2841233731 7310172001 7310172101

37 1 2 1 173 173

# find the maximum number of votes per ProductID

> head(ProductID\_votedtotal\_max)

ProductID\_votedtotal\_max <- tapply(df$Voted\_total, df$ProductId, max)

0006641040 141278509X 2734888454 2841233731 7310172001 7310172101

72 1 1 0 51 51

# find the total number of votes per ProductID

ProductID\_votedtotal\_sum <- tapply(df$Voted\_total, df$ProductId, sum)

> head(ProductID\_votedtotal\_sum)

0006641040 141278509X 2734888454 2841233731 7310172001 7310172101

125 1 1 0 211 211

# ----------------------------------------------------------------------------

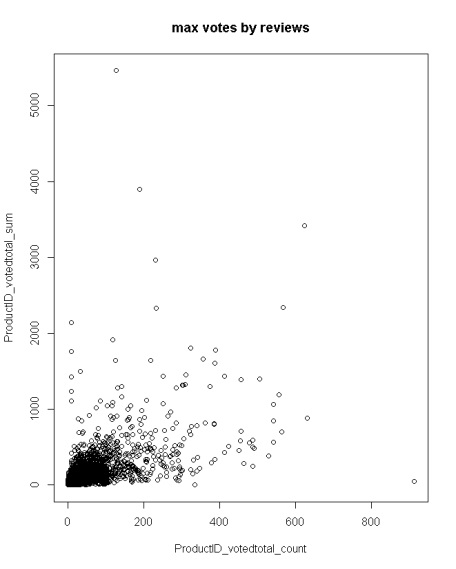
# PROBLEM 1H

# Make a scatterplot of max number of votes as a function of number of reviews.

# Is there a visible trend? If so, describe it.

# plot number of votes vy number of reviews

plot(ProductID\_votedtotal\_count, ProductID\_votedtotal\_sum, main="max votes by reviews")



ANALYSIS: The number of unique reviews tend to range roughly from 0 to 100 in the total number of reviews per ProductID. As the scatterplot depicts, there is a positive and increasing relationship between the number of reviews per ProductID and the total number of votes for the ProductID. This relationship is concentrated in the lower bounds of the range for these two attributes, but there are some outliers.

# ----------------------------------------------------------------------------

# PROBLEM 1J

# Histograms of the review counts and number of votes indicate that both variables are right-skewed

plotNormalHistogram(ProductID\_votedtotal\_count, xlim=c(0,200))

plotNormalHistogram(ProductID\_votedtotal\_sum)

plotNormalHistogram(ProductID\_votedtotal\_max)

# Subset the variables max.votes and number.of.reviews to only those values

# corresponding to products with 1 or more votes.

df\_one\_or\_more\_votes <- df[df$Voted\_total > 0, ]

ProductID\_votedtotal\_count\_pos <- tapply(df\_one\_or\_more\_votes$Voted\_total, df\_one\_or\_more\_votes$ProductId, length)

ProductID\_votedtotal\_max\_pos <- tapply(df\_one\_or\_more\_votes$Voted\_total, df\_one\_or\_more\_votes$ProductId, max)

# ----------------------------------------------------------------------------

# PROBLEM 1K

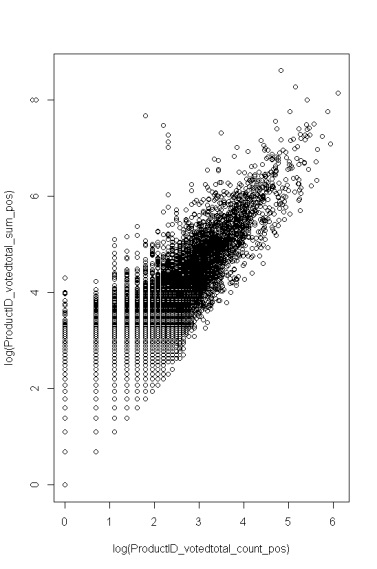
# Make a scatterplot of log(max.votes) as a function of log(number.of.reviews).

# Is there a visible trend? If so, describe it.

# Does this tell us anything about the relationship between the

# untransformed max.votes and number.of.reviews?

plot(log(ProductID\_votedtotal\_count\_pos), log(ProductID\_votedtotal\_max\_pos))



ANALYSIS: There seems to be a visible trend in the log(number of reviews) vs log(maximum number of votes) per ProductID in that there is a positive and increasing relationship between the two variables which correlates to the relationship that was found when the untransformed variables were compared. This indicates that these two variables are correlated logarithmically.