Math23c Final Project

MATH 23C PROJECT: AN INVESTIGATION INTO BOSTON HOUSING AND WEATHER

By Rudra Barua, Hope Ha, and Matti Tan

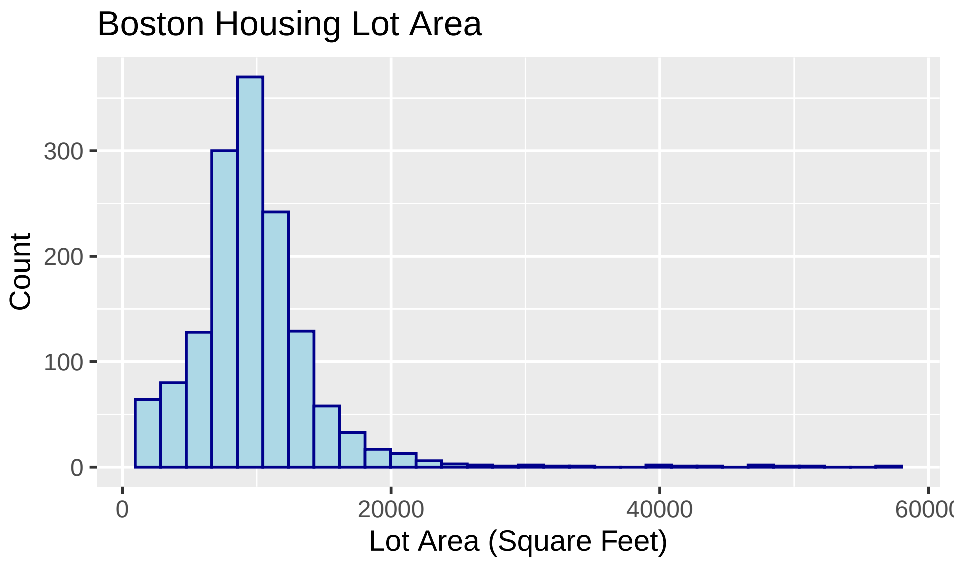
This is our investigation into a Boston Housing dataset and Boston weather dataset.

MODELLING THE DATA

We began by modeling the data through different graphical displays.

Histogram of Lot Area

```
print(LotAreaPlot + labs(title = "Boston Housing Lot Area"
                         ,y= "Count", x = "Lot Area (Square Feet)"))
```

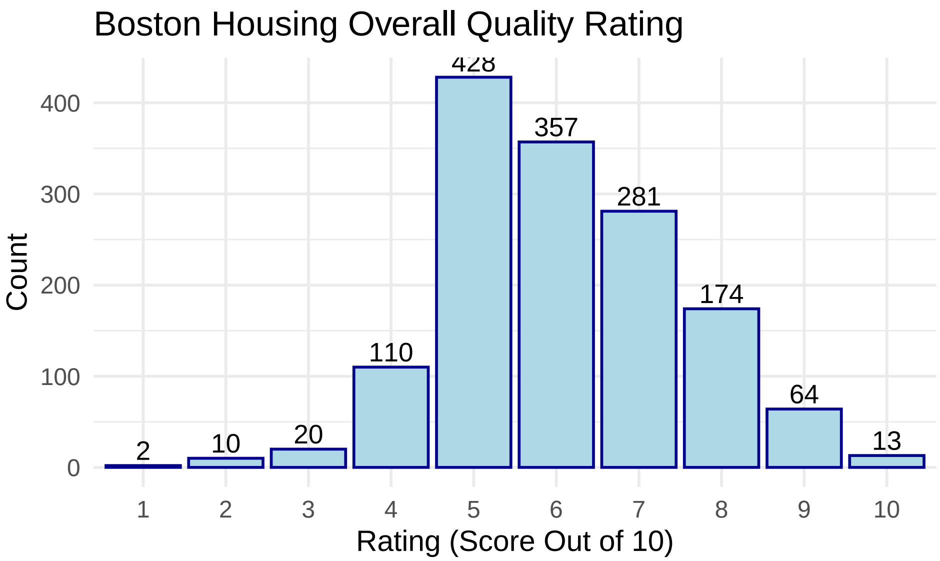


A histogram of the different lot areas displays a leftward skew. It appears to be that very few houses in Boston have lot areas greater than 20,000 square feet.

Barplot of Overall Quality

```
##    Var1 Freq
## 1     1    2
## 2     2   10
## 3     3   20
## 4     4  110
## 5      5  428
## 6      6  357
## 7      7  281
## 8      8  174
## 9      9   64
## 10    10   13
```

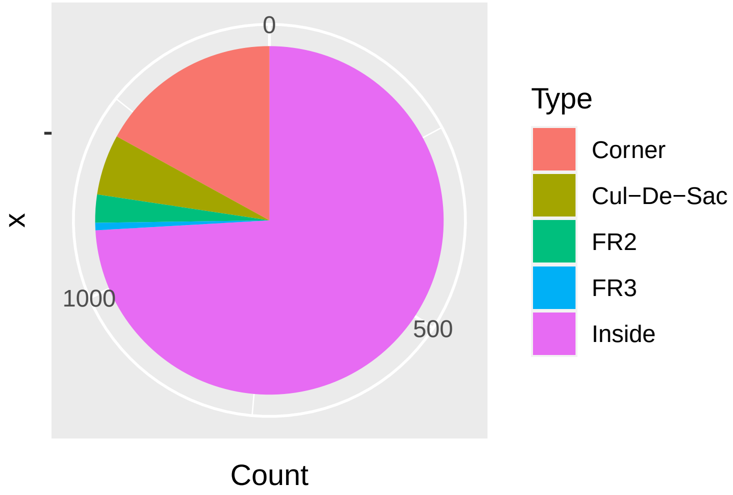
```
print(OverallQualityPlot + labs(title = "Boston Housing Overall Quality Rating",
                                y = "Count", x = "Rating (Score Out of 10)"))
```



The barplot displays how the most common rating of overall quality is 5 with 428 respondents giving this score. And, the least common raating is 1 with only 2 respondents giving this score.

Piechart of Lot Configuration

```
pie <- bp + coord_polar("y", start=0)
plot(pie)
```



This displays types of locations and shapes of the lot such as whether or not it is an inside lot (Inside), corner lot (Corner), cul-de-sac (CulDSac), a lot withfrontage on 2 sides of the property (FR2), or a lot with frontage on 3 sides of the property (FR3). The pie-chart above illustrates how a majority of Boston houses have "inside" lot configurations while house lots that have frontage in 3 sides of the property are the least common.

Having gotten some sense of the nature of our data from our graphical displays, we then performed a series of investigations to best analyze our data.

INVESTIGATION: How can we use novel statistics to examine the data for the surface area of open porches in Boston houses?

Let us take a more robust look at the data for the surface area of open porches in Boston houses. Since not all houses have open porches, we get a very wide spread of values that may lead us astray when we look at central tendency values.

Maximum and Minimum:

```
Porch <- boston$OpenPorchSF #extracting the data
max(Porch)
```

```
## [1] 742
```

```
min(Porch)
```

```
## [1] 0
```

As seen, we have a massive difference between the maximum and minimum values of porch surface area sizes.

```
length(which(Porch == 0))
```

```
## [1] 642
```

```
#we also have 642 houses that do not have open porches
642/1459 #this is about 44% of our data
```

```
## [1] 0.4400274
```

```
mean(Porch)
```

```
## [1] 48.31391
```

```
#Hence, this mean might not really capture the true average of Boston houses
#with porches because of the many outliers in this data.
```

Median:

```
median(Porch)
```

```
## [1] 28
```

This is the middle value in the data, and it is noticeably less than the mean.

Skewness:

```
skewness(Porch)
```

```
## [1] 2.682255
```

Our positive skewness indicates that the surface area of open porches in Boston houses is skewed right, which means that most data falls to the right side of the graph's peak, and generally (but not always), the mean is greater than the median (as is the case here).

Trimmed Mean: But, probably one of the best measures of central tendency for this data is the trimmed mean. By taking a trimmed mean, we remove a predetermined amount of the data (in this case, the lowest and highest values, outliers), and find an average for the remaining data observations. R has a built-in function for this:

```
mean(Porch, trim = 0.44) #trimming 44% from each side removes extreme values
```

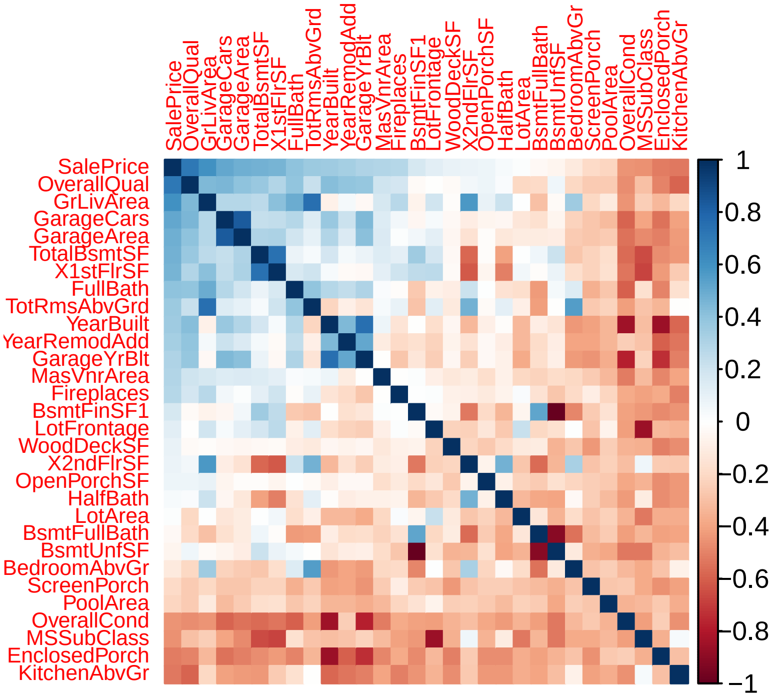
```
## [1] 26.82486
```

```
#This looks like a more reasonable mean surface area for the average Boston house porch
```

INVESTIGATION: What are the highest correlated variables? How might we model this?

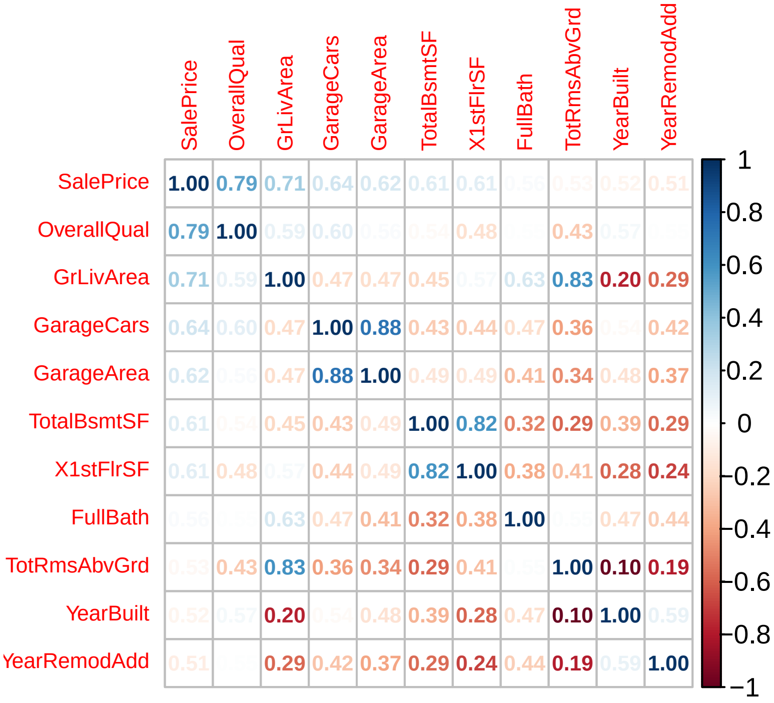
We would like to make a heatmap of the correlations of the numeric variables. We'll first do this for all correlations, ignoring ones that are less than 0.05.

```
corrplot(numcorrs, method="color",tl.cex = 0.7,number.cex=.7) # Pretty cool right!
```

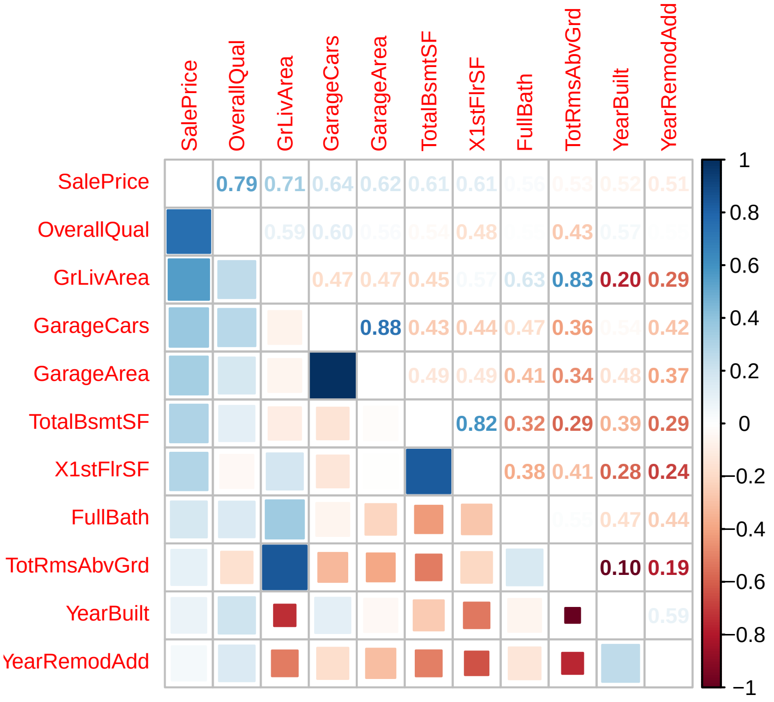


Now, lets make a correlation matrix for just the relationships that have a correlation above 0.5.

```
corrplot(numcorrs, method="number",tl.cex = 0.7,number.cex=.7) # Pretty cool right!
```



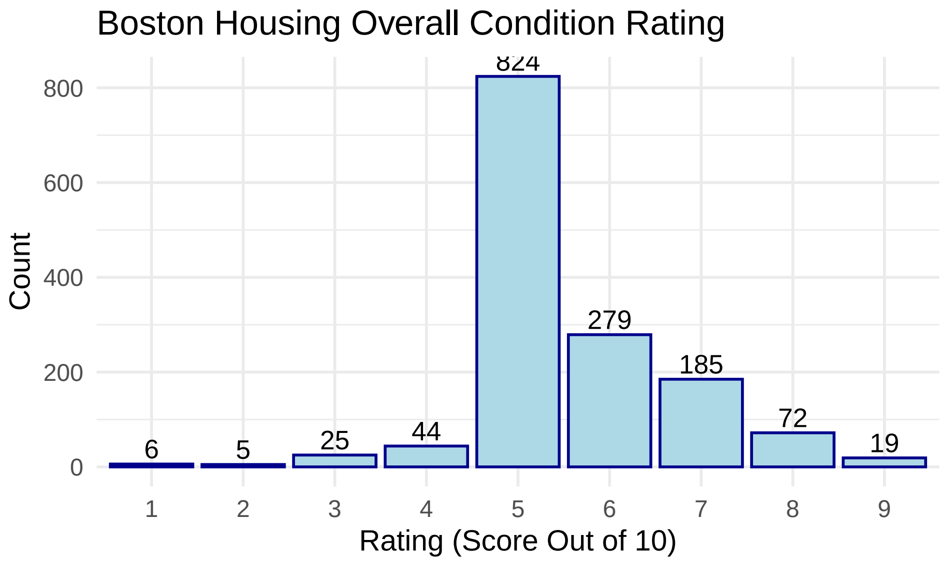
```
# A different style graph
corrplot.mixed(numcorrs, lower = "square", upper = "number", tl.cex = 0.7,cl.cex = .7, number.cex=.7,tl
```



INVESTIGATION: Is there a relationship between whether a house has central air conditioning and its overall condition?

Overall Condition measures the current state of the house in terms of wear and tear, physical comfort and present structural stability. It would be interesting to see if whether a house has central air conditioning has anything to do with it's overall condition. We can test this relationship using a permutation test. We first visualize the overall condition using a barplot:

```
print(OverallConditionPlot + labs(title = "Boston Housing Overall Condition Rating",
                                  y = "Count", x = "Rating (Score Out of 10)"))
```



Now, we can go ahead and conduct the actual permutation test:

```
sum(boston$CentralAir == "Y") #number of houses with central air conditioning
```

```
## [1] 1358
```

```
sum(boston$CentralAir == "N") #number of houses without central air conditioning
```

```
## [1] 101
```

```
# the observed averages of overall housing condition scores of houses with and without central air cond
CA.Yes.Avg <- sum(boston$OverallCond*(boston$CentralAir == "Y"))/sum(boston$CentralAir == "Y"); CA.Yes.
```

```
## [1] 5.590574
```

```
CA.No.Avg <- sum(boston$OverallCond*(boston$CentralAir == "N"))/sum(boston$CentralAir == "N"); CA.No.Avg
```

```
## [1] 5.059406
```

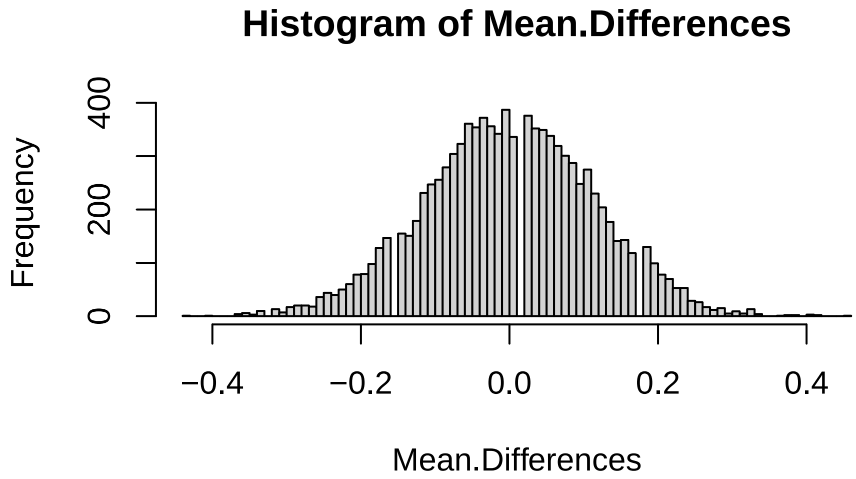
```
# observed difference
observed <- CA.Yes.Avg - CA.No.Avg; observed
```

```
## [1] 0.5311684
```

It appears that, on average, houses with central air conditioning have slightly higher scores.

We can then take 10000 permuted samples, and make a histogram of the mean differences in the permuted samples.

```
hist(Mean.Differences, breaks = "FD")
abline(v = observed, col = "red") #Distant from the most dense region
```



```
pvalue <- (sum(Mean.Differences >= observed)+1)/(N+1); pvalue
```

```
## [1] 9.999e-05
```

Since we have an incredibly small p-value that is certainly < 0.05, we then have a minute chance that we'd observe such a discrepancy in our data. Therefore, there is sufficient evidence to reject the null hypothesis that there is no significant difference in the average overall condition score of Boston houses between houses that have and that do not have central air conditioning, and there is a relationship between whether a house has central air conditioning and its overall condition.

INVESTIGATION: Is there a relationship between whether a house has central air conditioning and whether a house has a paved driveway?

We can conduct this investigation by using a contingency table, followed by a chi-square test.

In making our contingency table, we can do it manually and by using the table function. Manually, this looks like:

```
# houses with Central Åir and Paved Driveways.
CentralPave <- which(BostonLogical$CenAir&BostonLogical$PD);head(CentralPave)
```

```
## [1] 1 2 3 4 5 6
```

```
length(CentralPave) # there are 1258 counts.
```

```
## [1] 1258
```

```
# houses with Central Air and WITHOUT Paved Driveways.
CentralNotPave <- which(BostonLogical$CenAir& !BostonLogical$PD)
length(CentralNotPave) # there are 74 counts.
```

```
## [1] 74
```

```
# houses WITHOUT Central Air and with Paved Driveways.
NotCentralPave <- which(!BostonLogical$CenAir& BostonLogical$PD)
length(NotCentralPave) # there are 43 counts.
```

```
## [1] 43
```

```
# houses WITHOUT Central Air and WITHOUT Paved Driveways.
NotCentralNotPave <- which(!BostonLogical$CenAir& !BostonLogical$PD)
length(NotCentralNotPave) # there are 52 counts.
```

```
## [1] 52
```

Thus, we would expect our contingency table to look like:

```
                                   CENTRAL AIR
                                 FALSE      TRUE

                         FALSE     52         43
```

PAVED DRIVEWAY

```
                          TRUE     74       1258
```

We can check this with the table function, and indeed, using our logical columns and our cleaned dataset returns the same contingency table which matches with what we expected:

```
library("printr")
table(BostonLogical$CenAir, BostonLogical$PD); table(bostonCleaned$CentralAir, bostonCleaned$PavedDrive
```

/	FALSE	TRUE
FALSE	52	43
TRUE	74	1258

/	N	Y
N	52	43
Y	74	1258

```
detach('package:printr', unload = TRUE)
```

We now can conduct a chi-squared test, where our null hypothesis is that whether a house in Boston has Central Air is independent of whether the house has a Paved Driveway.

```
Observed <- table(BostonLogical$CenAir, BostonLogical$PD); Observed
```

```
##
##          FALSE TRUE
##   FALSE     52   43
##   TRUE      74 1258
```

```
Expected <- outer(rowSums(Observed), colSums(Observed))/sum(Observed); Expected
```

```
##             FALSE       TRUE
## FALSE   8.388227   86.61177
## TRUE  117.611773 1214.38823
```

```
# Manually:
Chi2 <-sum((Observed-Expected)^2/Expected); Chi2 # to calculate the chi-squared value
```

```
## [1] 266.4426
```

```
Pvalue<- pchisq(Chi2,1,lower.tail = FALSE); Pvalue # to find the p-value
```

```
## [1] 6.764544e-60
```

```
# built-in test confirms our result!
chisq.test(BostonLogical$CenAir, BostonLogical$PD)
```

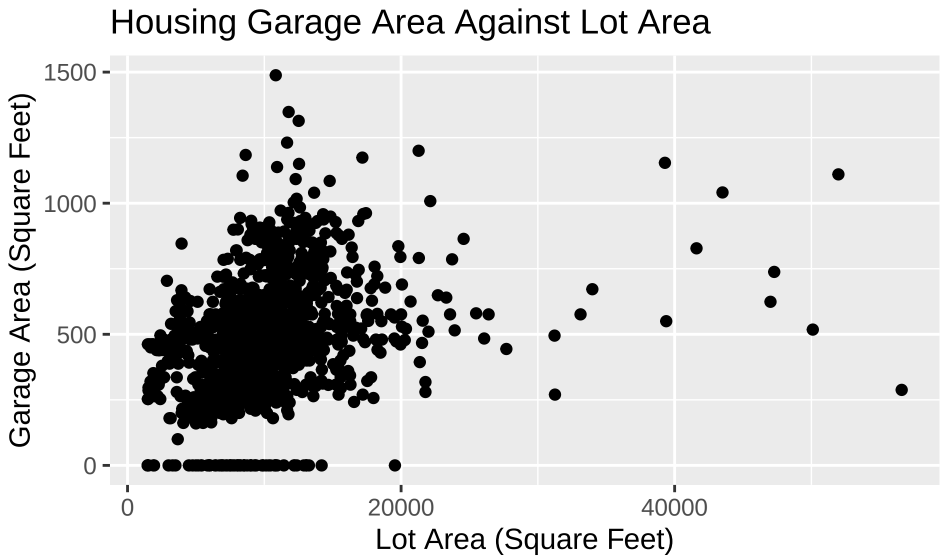
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  BostonLogical$CenAir and BostonLogical$PD
## X-squared = 260.37, df = 1, p-value < 2.2e-16
```

This returns that our p-value is 6.764544e-60, which means that the probability of this result occurring by chance is extremely low. As 6.764544e-60 is less than 0.01, we conclude that the result is significant, and as such, we can reject the null hypothesis that whether a house has central air conditioning is independent of whether a house has a paved driveway. We therefore conclude that whethera house has central air conditioning and whether a house has a paved driveway are dependent in this dataset (they share some sort of relationship).

INVESTIGATION: Does Lot Area predict Garage Area?

We shall investigate whether or not Lot Area predicts Garage Area using linear regression. This is interesting because it helps us understand how households allocate the area that they purchase. Hence, we are using Lot Area (in square feet) as the predictor for Garage Area (in square feet).

```
#Making a scatter plot of the data
LotvGaragePlot <- ggplot(boston, aes(x=LotArea, y=GarageArea)) +
  geom_point()

print(LotvGaragePlot + labs(title = "Housing Garage Area Against Lot Area"
                           , x = "Lot Area (Square Feet)",
                           y = "Garage Area (Square Feet)"))
```

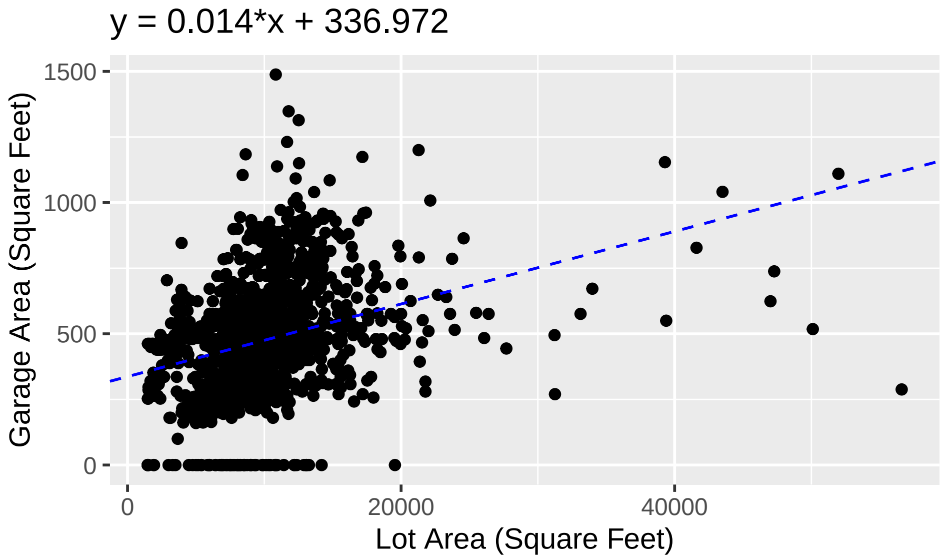


```
#Using the built-in R function
Boston.lm <- lm(boston$GarageArea ~ boston$LotArea, data = boston); Boston.lm
```

```
##
## Call:
## lm(formula = boston$GarageArea ~ boston$LotArea, data = boston)
##
## Coefficients:
##    (Intercept)  boston$LotArea
##      336.97230         0.01383
```

```
#Plotting the regression line
coeff=coefficients(Boston.lm)
#Regression Line Equation
eq = paste0("y = ", round(coeff[2],3), "*x + ", round(coeff[1],3))

#Adding to plot
print(LotvGaragePlot +
        labs(title = "Housing Garage Area Against Lot Area"
             , x = "Lot Area (Square Feet)", y = "Garage Area (Square Feet)")
```

```
    + geom_abline(intercept = 336.97230, slope = 0.01383, color="blue",
                  linetype="dashed")+ ggtitle(eq))
```



Hence, the line of best fit that we find through linear regression depicts how there might be a positive relationship/correlation between lot area and garage area. An increase in lot area might predict an increase in the garage area of the house. Intuitively, this is sensible since a larger lot area will allow for more land to be allocated for garages. Intriguingly, this might be indicative of how people choose to use the lot area they have.

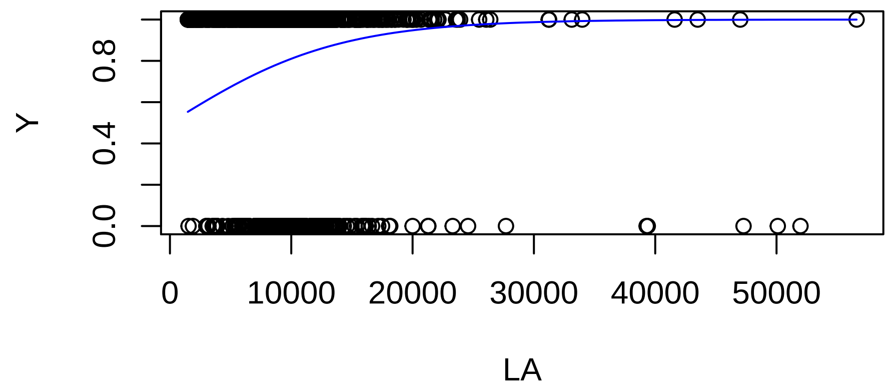
INVESTIGATION: Is the lot area of a house connected to the type of sale condition of Boston Houses?

We shall next investigate whether the lot area of a house is connected to the type of sale using logistic regression, which could be an indicator of the type of buyers for Boston houses. That is, we will use lot area to determine the type of sale condition of Boston houses; there are usually two types of sale condition: Normal (typical single buyer and seller) and Non-normal (Sold to a family, possible loan etc.)

```
library(stats4) #will need this library (install if necessary)
#In this revised data set, the sale type entries were replaced with binary
#values 1 and 0 from "Normal" and the other non-normal entries, respectively.

#Now we can extract the  as a Bernoulli random variable:
Y <- boston$SaleCondition

#To perform logistic regression, we will assume that
#p = exp(alpha x+beta)/(1 + exp(alpha x+beta))
#where 0 =< p =< 1

#First, we shall model the probability of a sale condition
# as a function of lot area
#Here, the predictor is the lot area column
LA <- boston$LotArea

#Creating the log-likelohood function for lot frontage as the predictor
MLL<- function(alpha, beta) {
  -sum( log( exp(alpha+beta*LA)/(1+exp(alpha+beta*LA)) )*Y
        + log(1/(1+exp(alpha+beta*LA)))*(1-Y) )
}

results<-mle(MLL, start = list(alpha = 0, beta = 0))#initial guess
results@coef#from this we know that...
```

```
##        alpha         beta
## 1.554842e-08 1.451108e-04
```

```
#alpha = 1.554842e-08
#beta = 1.451108e-04
plot(LA,Y)
curve( exp(results@coef[1]+results@coef[2]*x)/ (1+exp(results@coef[1]+results@coef[2]*x)),col = "blue",
```



The curve found through logistic regression appears to be a reasonable approximation of how lot area and sale type are related due to the points in the upper portion of the graph being crossed by the curve. Hence, there appears to be a positive relationship between normal type sales and lot area. However, without an appropriate statistical test and as the lower points are not captured by the curve, it is not possible to draw a definite conclusion.

INVESTIGATION: Is there a statistically significant difference between the average sale price in a house in a cul-de-sac versus a house on a corner?

To answer this, we can use both simulation and classical means. We begin by using simulation with a permutation test.

```
# Get the average sale price of CulDSac houses
CDavg <- mean(CDprice); CDavg
```

```
## [1] 223854.6
```

```
COprice <- config$SalePrice[idxCO]
# Get the average sale price of Corner houses
COavg <- mean(COprice); COavg
```

```
## [1] 181623.4
```

```
# Take the difference of the two to see which has the higher number average sale price
obsdiff <- CDavg - COavg; obsdiff
```

```
## [1] 42231.19
```

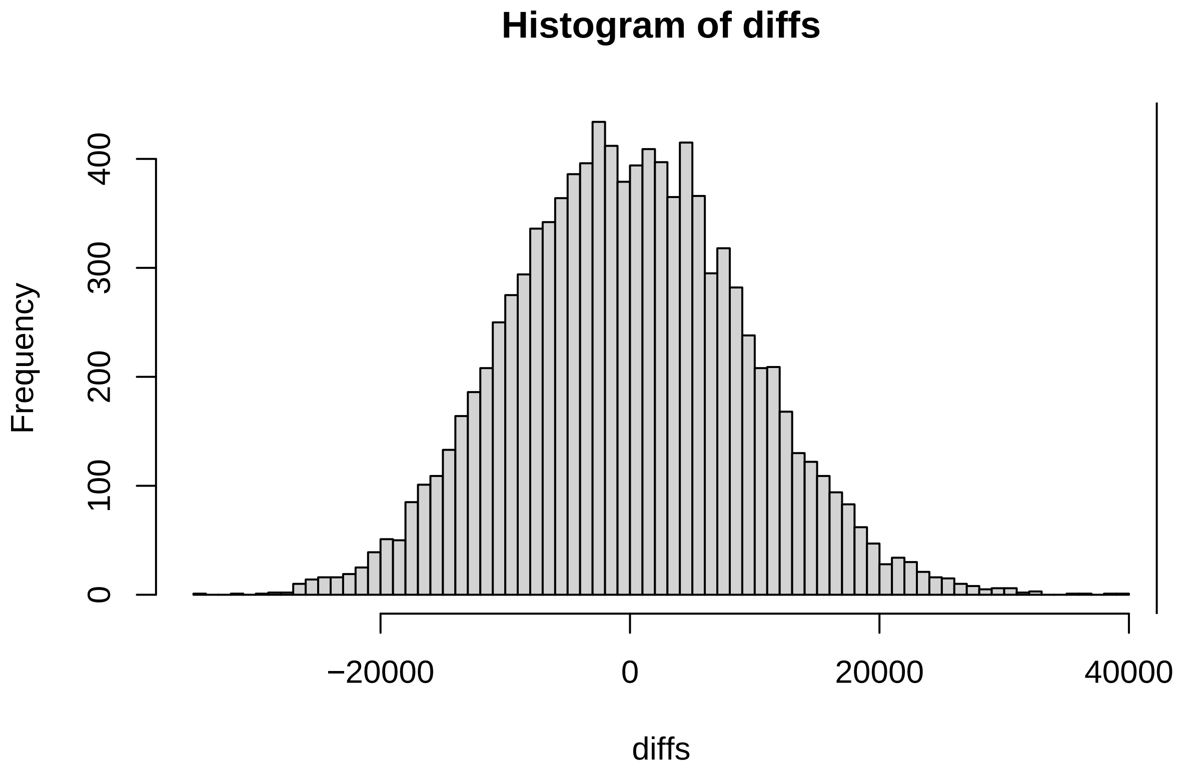
```
# This implies that on average, houses on cul-de-sacs sell at a higher sale price
# than houses on corners
```

We can then conduct our permutation test.

```
# Repeat 10000 times for our permutation test
N <- 10000
diffs <- numeric(N)
for (i in 1:N){
  type <- sample(config$LotConfig) # Permuted lot config column
  CDavg <- sum(config$SalePrice*(type=="CulDSac"))/sum(type=="CulDSac")
  COavg <- sum(config$SalePrice*(type=="Corner"))/sum(type=="Corner")
  diffs[i] <- CDavg - COavg   # As likely to be negative or positive
}
mean(diffs) # Should be close to zero
```

```
## [1] -101.5309
```

```
hist(diffs, breaks = "FD")
# Now display the observed difference on the histogram
abline(v=obsdiff)
```



```
pvalue <- (sum(diffs >= obsdiff)+1)/(N+1); pvalue
```

```
## [1] 9.999e-05
```

The probability (the P value, 9.999e-05) that a difference this large could have arisen with a random subset is 0.019998%.Our data provides sufficient evidence against the null hypothesis of there being no difference in the average sale price of a house located in a cul-de-sac over a house located on a corner. Instead it provides sufficient evidence for the hypothesis that there is a difference in the average sale price of a house located in a cul-de-sac over a house located on a corner. We can asses that the average sale price of a house on a cul-de-sac in Boston is greater than the average sale price of a house on a corner in Boston.

We can now use classical means and compare.

```
# Now let us look at a two-sample t-test
t.test(config$SalePrice[idxCD], config$SalePrice[idxCO], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  config$SalePrice[idxCD] and config$SalePrice[idxCO]
## t = 4.048, df = 355, p-value = 6.344e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21713.77 62748.61
## sample estimates:
```

```
## mean of x mean of y
##  223854.6  181623.4
```

This also gives us a p-value of 6.344e-05, which is very comparable to our earlier p-value, and confirms our earlier results.

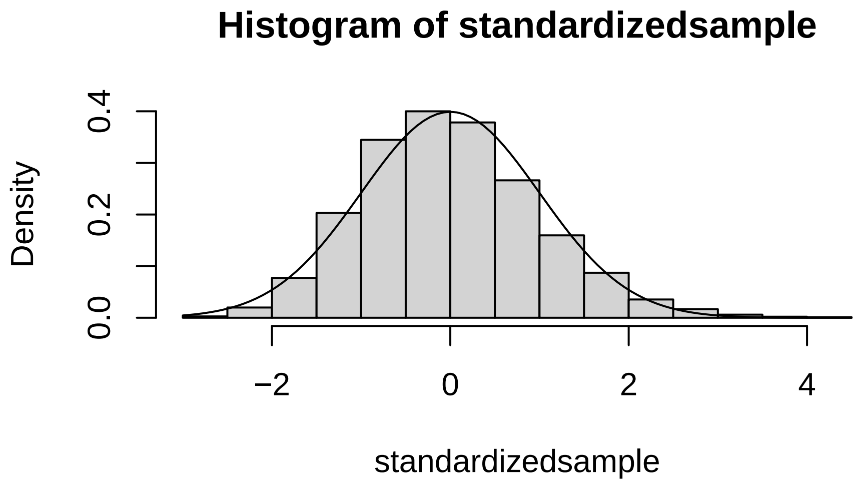
Notably, although they are comparable, our permutation test results are more reliable because our two-sample t-test particularly relies on the assumption that our data follows a normal distribution, while the permutation test has no such assumptions or requirements.

INVESTIGATION: Can we check whether a sampling distribution of the above grade (ground) living area in square feet in Boston is standard normal?

If we let GrLivArea, the above grade (ground) living area square feet, be equal to random variable Y, it might not be unreasonable to assume that the random variable $X = \frac{(Y-\mu)}{\sigma}$ has a standard normal distribution, especially given the large number of data points we have. According to the Central Limit Theorem, if we take our population with mean μ and standard deviation σ and we take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed. Thus, we would theoretically expect our random variable to have a standard normal distribution.

We want to perform a simulation by taking samples of six houses at a time, to see whether this sampling distribution behaves as if drawn from a population with a normal distribution.

```
GrLivArea <- boston$GrLivArea
mu <- mean(GrLivArea)
sigma <- sd(GrLivArea)
N <- 10000
n <- 6
get.sample <- function() sample(GrLivArea, n) # defining our sampling function
standardizedsample <- numeric(N)
for(i in 1:N) {
  x <- get.sample()
  standardizedsample[i] <- (mean(x) - mu)/(sigma / sqrt(n))
}
hist(standardizedsample, prob = TRUE) # histogram
curve(dnorm(x), add = TRUE) # normal distribution curve
```



Indeed, a standard normal curve appears to fit our histogram well, and so our variable X does seem to be distributed standard normal.

INVESTIGATION: What range of house sale prices captures the mean 95% of the time?

We can use a student-t confidence interval to answer this question.

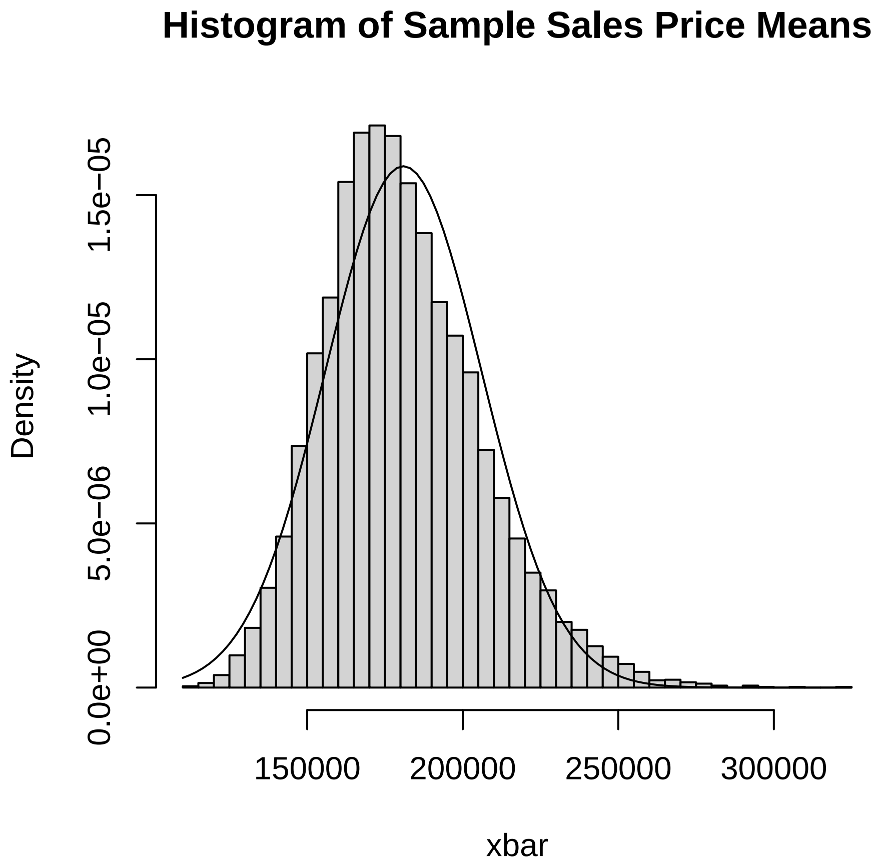
```r
# Let us make a histogram of the means for 10000 samples
N <- 10^4 # Trials
n <- 10 # Sample size
mu <- mean(boston2$SalePrice); mu # Mean: 180921.2
```

```
## [1] 180921.2
```

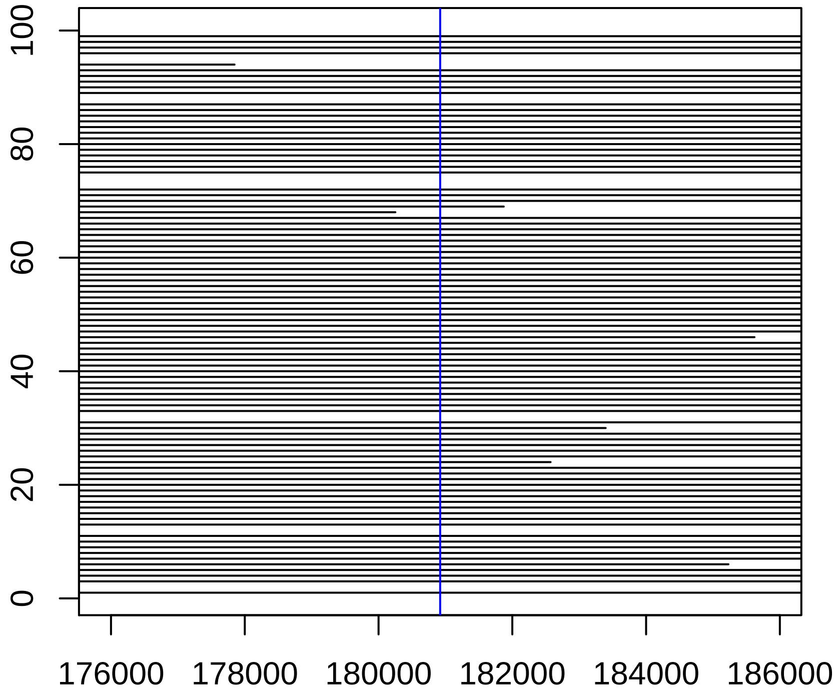
```r
sigma <- sd(boston2$SalePrice); sigma # Standard dev: 79442.5
```

```
## [1] 79442.5
```

```r
xbar <- numeric(N) # To store our results

# Let us generate our samples
for (i in 1:N) {
  samp <- sample(boston2$SalePrice, n) # Take size n sample
  xbar[i] <- mean(samp) # Store the sample mean
}
# Now lets compare the samples with the corresponding normal distribution.
hist(xbar, probability = TRUE, main = "Histogram of Sample Sales Price Means", breaks = "FD")

# Overlay normal density curve
curve(dnorm(x, mu, sigma/sqrt(n)), add = TRUE) # It fits pretty well
```



```r
# Now let us construct the student t 95% confidence interval
# This is a modification of Paul's 8D Script
# Initialize the counts for missing the lower/upper points of confidence interval
missedL <- 0; missedU <- 0
# Create a good sized plot
plot(x = c(mu - 5000, mu + 5000), y = c(1,100), type = "n", xlab = "", ylab = "")
# Lets take 1000 samples
N <- 1000
counter <- 0 # Set counter to 0
# Run the loop
for (i in 1:N) {
  x <- sample(boston2$SalePrice, n) # Take samples size of 10
  L <- mean(x) + qt(0.025, n - 1) * sd(x)/sqrt(n) # Lower endpoint
  U <- mean(x) + qt(0.975, n - 1) * sd(x)/sqrt(n) # Upper endpoint
  # Increment the value is missing the upper or lower endpoint
  if (mu < L) missedL <- missedL + 1
  if (mu > U) missedU <- missedU + 1
  if (L < mu && U > mu) counter <- counter + 1
  if (i <= 100) segments(L, i, U, i) # Plot the results
}

# Plot a vertical line for the true mean
abline(v = mu, col = "blue")
```



```
# What fraction of the time did the interval include the true mean?
(N - missedL - missedU)/N
```

```
## [1] 0.914
```

```
# What fraction of the time did it miss to the left?
missedL/N
```

```
## [1] 0.005
```

```
# What fraction of the time did it miss to the right?
missedU/N
```

```
## [1] 0.081
```

```
# This confidence interval is not to be trusted because the errors are not symmetrical.

# We can use the built-in t-test function to confirm:
t.test(boston2$SalePrice, conf.level=0.95)$conf.int
```

```
## [1] 176842.8 184999.6
## attr(,"conf.level")
## [1] 0.95
```

INVESTIGATION: Does the average wind in Boston from 2013-2018 follow a Gamma Distribution?

We began by investigating whether the average wind in Boston from 2013-2018 follows a Gamma distribution. As we saw in Paul's gamma fitter app, the average wind speed at the Carleton Turbine can be fitted using a gamma distribution quite well.

As we are also exploring Boston Housing, we'd also like to explore Boston temperature and see whether the average wind speed in Boston also can be well-approximated using a gamma distribution.

We used Paul's function to calculate the negative log-likelihood to find our parameters for a gamma distribution based on the data, and we got:

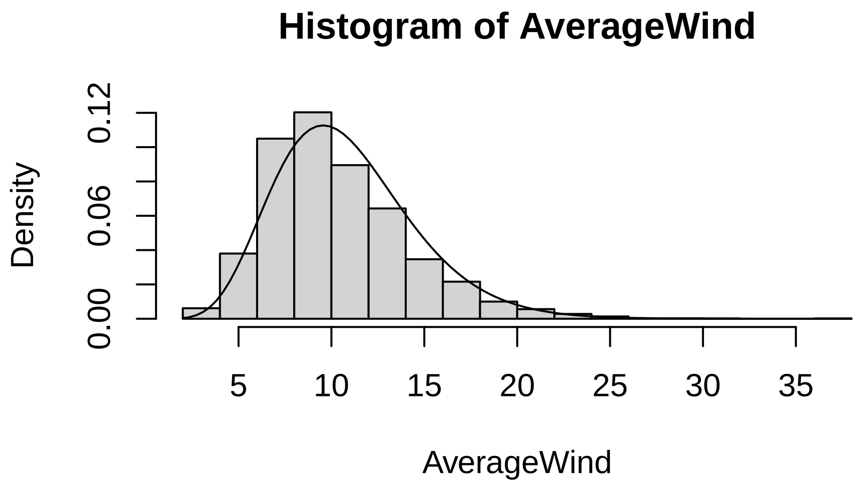
```
MLL<- function(shape, rate) {
  -sum(log(dgamma(AverageWind,shape,rate)))
}
results<-mle(MLL, start = list(shape = xshape, rate  = xrate)) #an initial guess is required
shape <- results@coef[1]; rate <- results@coef[2]
shape; rate
```

```
##   shape
## 8.44354
```

```
##      rate
## 0.7792171
```

We then laid a gamma distribution curve over the data with these parameters to see whether it would be a good fit.

```
hist(AverageWind, prob=TRUE)
curve(dgamma(x, shape, rate=rate), add = TRUE)
```



It indeed seems to fit well!

We can also check the mean and variance of the average wind speed vector, and see if it matches with the calculated mean and variance based on our gamma distribution:

```
mean(AverageWind); var(AverageWind) # the true mean and variance
```

```
## [1] 10.83596
```

```
## [1] 14.74496
```

```
round(shape/rate,4); round(shape/rate^2,6) # mean and variance based on gamma distribution
```

```
##   shape
## 10.8359
```

```
##    shape
## 13.90617
```

```
# indeed, they are very close!
```

We then conducted a chi-square test to see if we could see if the average wind speed truly had a gamma distribution.

We start with binning our observed values into 10 categories, and calculating our expected values.

```
bins <- qgamma(0.1 * (0:10), shape = 8.44354, rate = 0.7792171) # using our found parameters
observed <- as.data.frame(table(cut(AverageWind, bins, labels = FALSE)))$Freq
observed
```

```
##  [1] 331 336 451 449 453 360 311 247 376 435
```

```
# to calculate our expected values:
expected <- sum(observed)/10; expected
```

```
## [1] 374.9
```

We can then calculate the chi-squared value and p-value:

```
Chi2 <-sum((observed-expected)^2/expected); Chi2
```

```
## [1] 120.2958
```

```
Pvalue<- pchisq(Chi2,8,lower.tail = FALSE); Pvalue
```

```
## [1] 2.880328e-22
```

So, as our p-value is 2.880328e-22, which is less than 0.05, we have found that our result is statistically significant. As such, we reject the null hypothesis that our data follows a gamma distribution, despite our earlier analysis that suggested the average windspeed DID follow a gamma distribution.