HW3_Roser

Question 1: What is the effect of cut quality on diamond price?

The effect of each cut on diamond price is as follows: From a Fair to Good cut diamond, the price decreases by an average of 429 dollars. From a Fair to a Very Good cut diamond, the price decreases by an average of 377 dollars. From a Fair to Ideal cut diamond, the price decreased by an average of 901 dollars. From a Fair to Premium cut diamond, the price increased by an average 225 dollars. You only care about the effect of cut on price of diamonds, the only cut worthwhile is premium. There's a signficant difference in the effect of cut on diamond price because none of the confidence intervals cross zero.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```r
library(ggplotgui)
```

```
## Warning: package 'ggplotgui' was built under R version 3.4.4
```

```r
#ggplot_shiny(diamond)

diamond<-read.csv("diamond.csv")

head(diamond)
```

```
##   price       cut carat
## 1   326     Ideal  0.23
## 2   326   Premium  0.21
## 3   327      Good  0.23
## 4   334   Premium  0.29
## 5   335      Good  0.31
## 6   336 Very Good  0.24
```

```r
str(diamond)
```

```
## 'data.frame':    53940 obs. of  3 variables:
##  $ price: int  326 326 327 334 335 336 336 337 337 338 ...
##  $ cut  : Factor w/ 5 levels "Fair","Good",..: 3 4 2 4 2 5 5 5 1 5 ...
##  $ carat: num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
```

```r
price_over_carat<-diamond$price/diamond$carat

gem_mod<-glm(price ~ cut, data = diamond, family = "poisson")
coef(gem_mod)
```

```
##  (Intercept)      cutGood     cutIdeal   cutPremium cutVery Good
##    8.3799424   -0.1038367   -0.2316292    0.0504411   -0.0904632
```

```
exp(8.3799)  # $4358 is baseline for fair cut diamonds,
```

```
## [1] 4358.573
```

```
exp(8.3799-0.1038)-exp(8.3799) #fair compared to good
```

```
## [1] -429.7311
```

```
exp(8.3799-0.2316)-exp(8.3799) #fair compared to ideal
```

```
## [1] -901.0767
```

```
exp(8.3799+0.0504)-exp(8.3799) #fair compared to premium
```

```
## [1] 225.302
```

```
exp(8.3799-0.0906)-exp(8.3799) #fair compared to very good
```

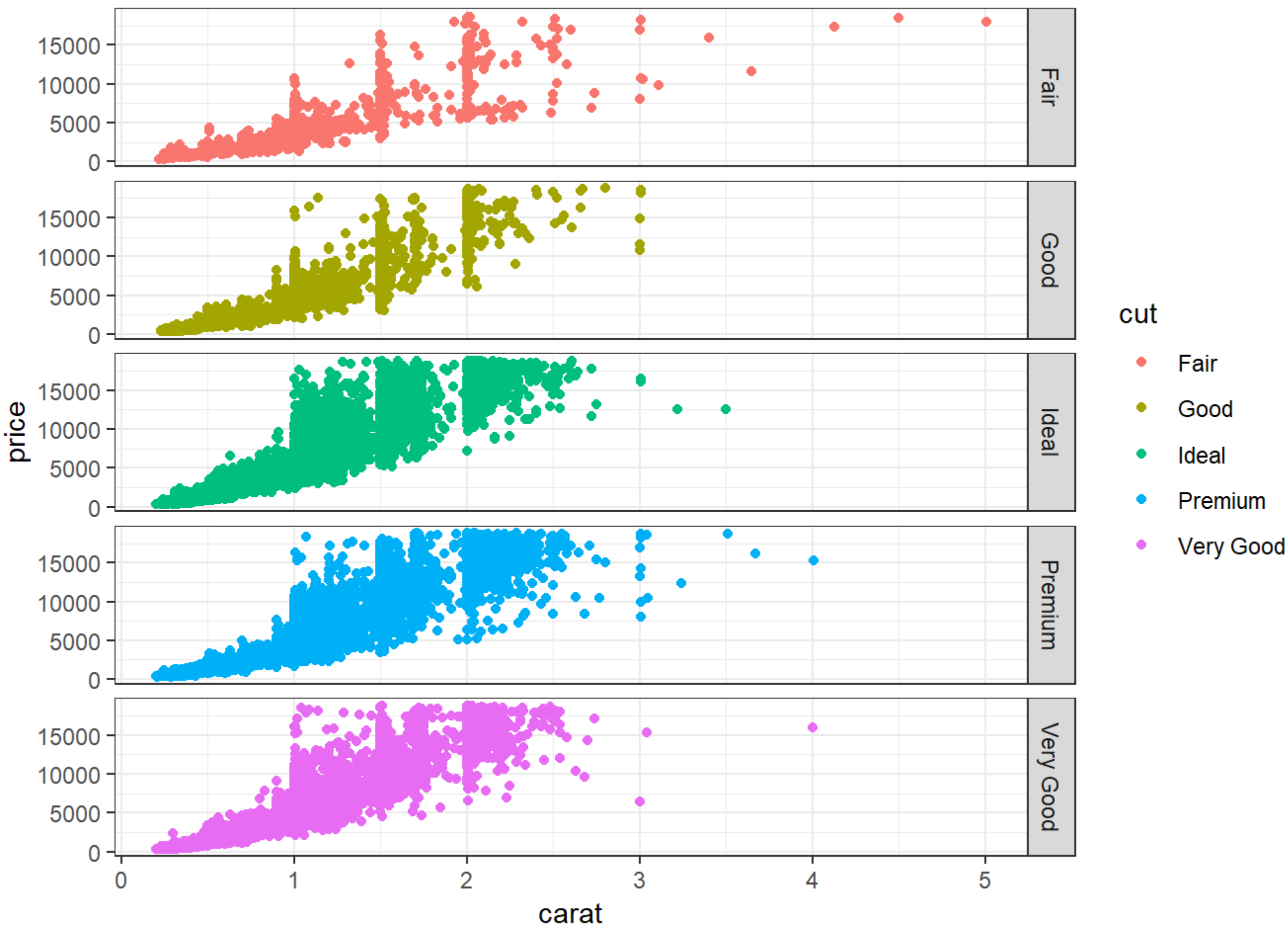
```
## [1] -377.5266
```

```
confint(gem_mod)  # all effects are significant
```

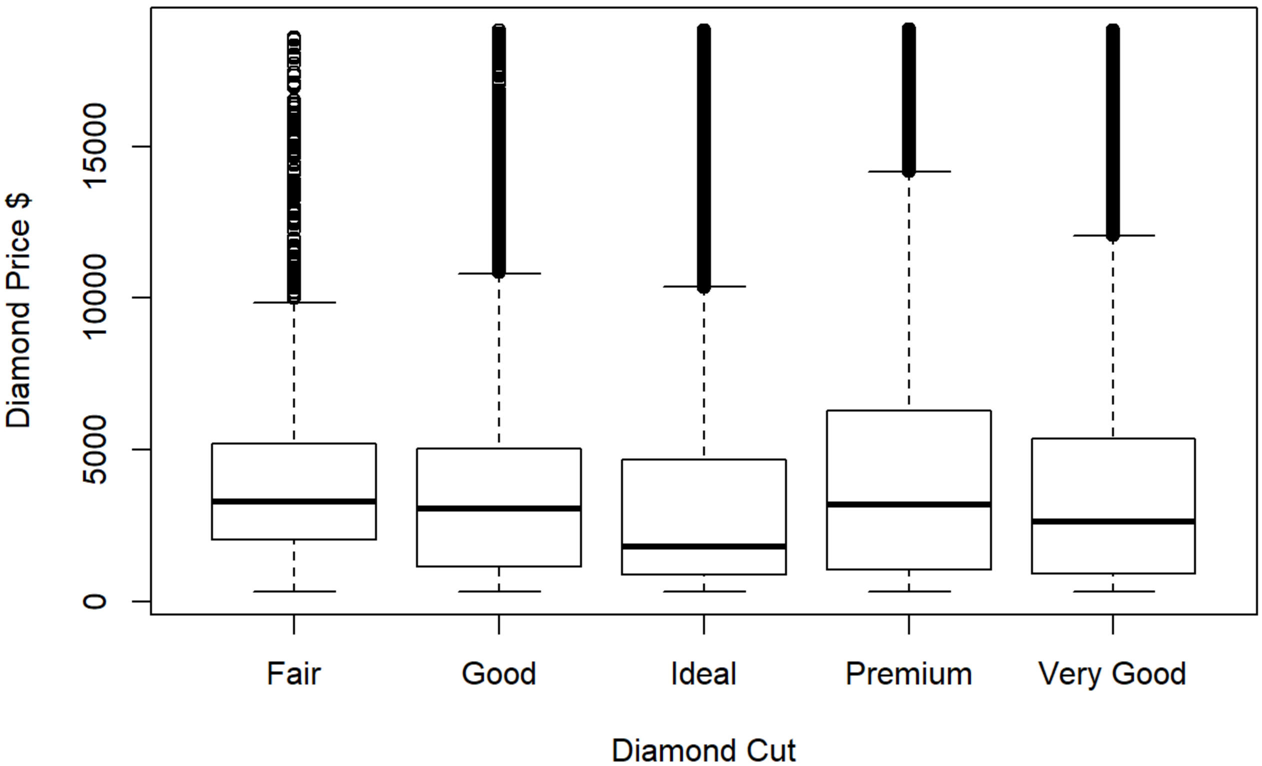
```
## Waiting for profiling to be done...
```

```
##                     2.5 %      97.5 %
## (Intercept)   8.37920242  8.38068216
## cutGood      -0.10470072 -0.10297248
## cutIdeal     -0.23240302 -0.23085517
## cutPremium    0.04966133  0.05122103
## cutVery Good -0.09125511 -0.08967112
```

```
scatter_diamond<- ggplot(diamond, aes(x = carat, y = price, colour = cut)) +
  geom_point() +
  facet_grid( cut ~ . ) +
  theme_bw()
scatter_diamond
```




```
boxplot(price~cut, data = diamond, xlab = "Diamond Cut", ylab = "Diamond Price $")
```




```
#boxplot(price_over_carat~cut, data = diamond, xlab = "Diamond Cut", ylab = "Diamond Price$/Carat")
```

Question 2: Does education have an impact on contraception use?

Women who have “high” education are on average 2% more likely to use contraception than women with “low” education. However, the CI intervals cross 0 so there’s not a significant effect of education level on contraception use.

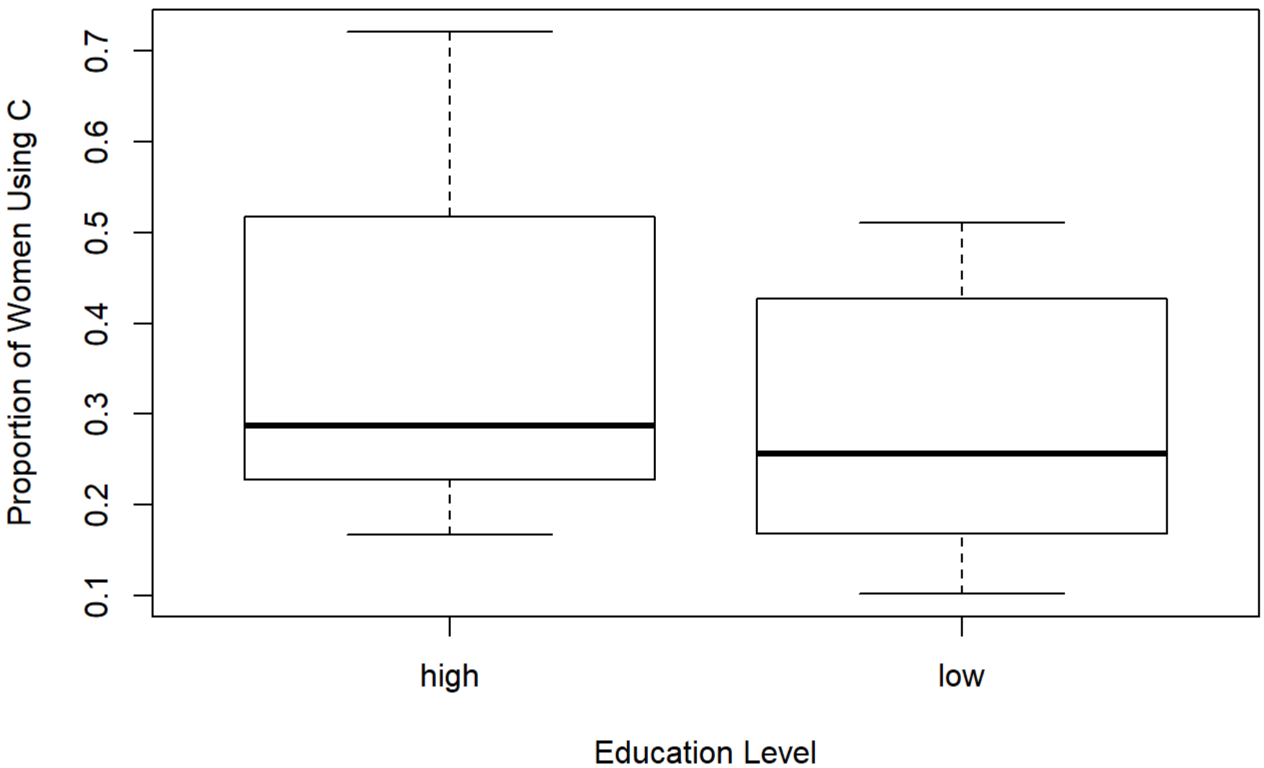
```
cuse<-read.csv("contraception.csv")

head(cuse)
```

```
##     age education notUsing using Total
## 1   <25       low       53     6    59
## 2   <25       low       10     4    14
## 3   <25      high      212    52   264
## 4   <25      high       50    10    60
## 5 25-29       low       60    14    74
## 6 25-29       low       19    10    29
```

```
str(cuse)
```

```
## 'data.frame':    16 obs. of  5 variables:
##  $ age      : Factor w/ 4 levels "<25","25-29",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ education: Factor w/ 2 levels "high","low": 2 2 1 1 2 2 1 1 2 2 ...
##  $ notUsing : int  53 10 212 50 60 19 155 65 112 77 ...
##  $ using    : int  6 4 52 10 14 10 54 27 33 80 ...
##  $ Total    : int  59 14 264 60 74 29 209 92 145 157 ...
```

```
prop_using<-cuse$using/cuse$Total
boxplot(prop_using~education, data = cuse, xlab = "Education Level", ylab = "Proportion of Women Using C")
  #visualize data
```




```
use_success<-cbind(cuse$using, cuse$notUsing)    #make response variable with both outcomes
use_mod<-glm(use_success~cuse$education, family = "binomial")

coef(use_mod)
```

```
##       (Intercept) cuse$educationlow
##       -0.81020374        0.09248529
```

```
plogis(-0.8102 + 0.0924)-plogis(-0.8102)  #difference between total women who use contraception
```

```
## [1] 0.02002974
```

```
confint(use_mod)
```

```
## Waiting for profiling to be done...
```

```
##                        2.5 %     97.5 %
## (Intercept)       -0.9460962 -0.6766394
## cuse$educationlow -0.1239481  0.3078275
```

Question 3: Hurricanes and Himmicanes

Based on my analysis of the deadliness of hurricanes vs himmicanes, I observed that there are on average 9 more deaths in hurricanes with female names than in hurricanes with male names. The 95% CI interval does not cross zero, from which we can infer a significant effect of himmicanes vs hurricanes and their respective deadliness.

I'm unsure as to how Jung et al could have provided more confidence in their analyses because I ran the same data through a negative binomial which showed no significant effect on average deaths between himmicanes and hurricanes. I believe this is because the poisson distribution which was used first does not accurately represent the large variance of the hurricane dataset. The negative binomial distribution is a better fit from this data because the variance is much greater than the mean.

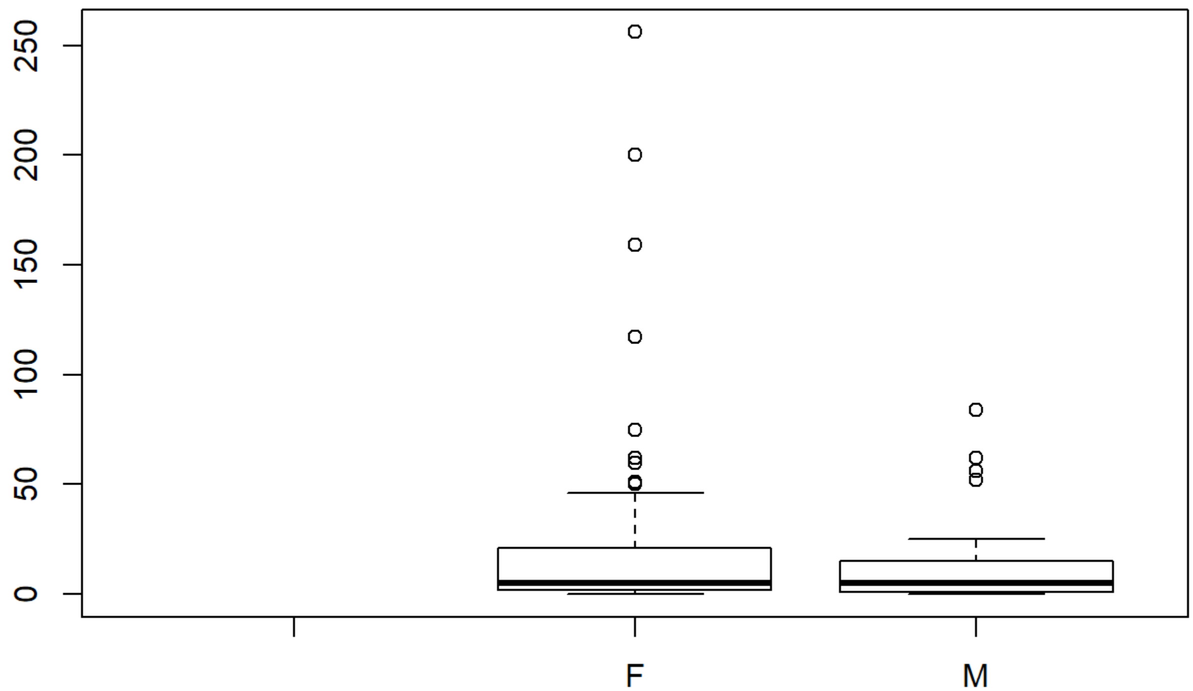
```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.4.4
```

```
storm<-read.csv("Hurricane Dataset.csv")

head(storm)
```

```
##   Year     Name  MasFem MinPressure_before Minpressure_Updated.2014
## 1 1950     Easy 6.77778                958                      960
## 2 1950     King 1.38889                955                      955
## 3 1952     Able 3.83333                985                      985
## 4 1953  Barbara 9.83333                987                      987
## 5 1953 Florence 8.33333                985                      985
## 6 1954    Carol 8.11111                960                      960
##   Gender_MF Category alldeaths  NDAM Elapsed.Yrs Source  ZMasFem
## 1         F        3         2  1590          63    MWR -0.00094
## 2         M        3         4  5350          63    MWR -1.67076
## 3         M        1         3   150          61    MWR -0.91331
## 4         F        1         1    58          60    MWR  0.94587
## 5         F        1         0    15          60    MWR  0.48108
## 6         F        3        60 19321          59    MWR  0.41222
##   ZMinPressure_A    ZNDAM
## 1       -0.35636 -0.43913
## 2       -0.51125 -0.14843
## 3        1.03765 -0.55047
## 4        1.14091 -0.55758
## 5        1.03765 -0.56090
## 6       -0.25310  0.93174
```

```
boxplot(alldeaths~Gender_MF, data = storm)    #note extreme outlines in Female named hurricanes
```




```
storm_mod<-glm(alldeaths~Gender_MF, data = storm, family = "poisson")
coef(storm_mod)
```

```
## (Intercept)  Gender_MFM
##   3.1679220  -0.5123354
```

```
exp(3.167-0.5123)-exp(3.167)  #on average there's 9 more deaths in hurricanes than himmicanes
```

```
## [1] -9.51545
```

```
confint(storm_mod)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  3.1164152  3.2185581
## Gender_MFM  -0.6211542 -0.4056501
```

```
#testing with negative rbinom
storm_mod2<-glm.nb(alldeaths~Gender_MF, data = storm)
coef(storm_mod2)
```

```
## (Intercept)  Gender_MFM
##   3.1679220  -0.5123354
```

```
confint(storm_mod2)  #shows the M/F is not significant effect on deaths--> poisson is nota good choice bc v
ariance is not well represented for this data set.
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept)  2.816448 3.5640722
## Gender_MFM  -1.149166 0.1720959
```

Question 4: Dataset from Our Own Research: Shrub Counts

The three shrub communities included are located at different elevations in Reynolds Creek Experimental Watershed; from low to high: Wyoming Big Sage, Low Sage, and Mountain Big Sage (2500ft to 7000ft). Site location has a significant effect on the density of shrubs; on average there's 10 more shrubs at a LOS plot than an MBS plot and 14 more at a LOS plot than WBS. In 2016, field crews also completed destructive above ground biomass sampling at each of these three sites. Site location also has a significant effect on the biomass of collected shrubs; we observe that on average, shrubs in LOS plots have the least biomass compared to WBS and MBS.

```
shrub<-read.csv("shrub_edit.csv")
head(shrub)
```

```
##   Site Plot      Date Recorder Observer1 Observer2 Plot_100m Plot_10m
## 1 wbs1    1 31-May-16     cami    jordan      alex        ne       ne
## 2 wbs1    1 31-May-16     cami    jordan      alex        ne       ne
## 3 wbs1    1 31-May-16     cami    jordan      alex        ne       ne
## 4 wbs1    1 31-May-16     cami    jordan      alex        ne       sw
## 5 wbs1    1 31-May-16     cami    jordan      alex        nw       ne
## 6 wbs1    1 31-May-16     cami    jordan      alex        nw       ne
##   Species Count Density Location
## 1   arar8     6     0.6      WBS
## 2  artrw8     1     0.1      WBS
## 3   chvi8     3     0.3      WBS
## 4   arar8     1     0.1      WBS
## 5   arar8     7     0.7      WBS
## 6  artrw8     3     0.3      WBS
```

```
str(shrub)
```

```
## 'data.frame':    355 obs. of  12 variables:
##  $ Site     : Factor w/ 9 levels "LOS1","LOS2",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ Plot     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date     : Factor w/ 17 levels "1-Jun-16","11-Jul-16",..: 15 15 15 15 15 15 15 15 15 15 ...
##  $ Recorder : Factor w/ 6 levels "alex","Alex",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Observer1: Factor w/ 5 levels "Alex,Cami","cami",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Observer2: Factor w/ 6 levels "","alex","Alex",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Plot_100m: Factor w/ 4 levels "ne","nw","se",..: 1 1 1 1 2 2 2 2 2 2 ...
##  $ Plot_10m : Factor w/ 4 levels "ne","nw","se",..: 1 1 1 4 1 1 1 2 2 2 ...
##  $ Species  : Factor w/ 10 levels "arar","arar8",..: 2 4 5 2 2 4 10 2 4 10 ...
##  $ Count    : int  6 1 3 1 7 3 1 3 3 1 ...
##  $ Density  : num  0.6 0.1 0.3 0.1 0.7 0.3 0.1 0.3 0.3 0.1 ...
##  $ Location : Factor w/ 3 levels "LOS","MBS","WBS": 3 3 3 3 3 3 3 3 3 3 ...
```

```
shrub_mod<-glm(Count~Location, data = shrub, family = "poisson")
coef(shrub_mod)
```

```
## (Intercept) LocationMBS LocationWBS
##   2.8873146  -0.7701327  -1.5061353
```

```
exp(2.887-0.7701)-exp(2.887)   #comparison between LOS and MBS    On average there's 10 more shrubs in LOS
 plots than MBS plots
```

```
## [1] -9.63406
```

```
exp(2.887-1.5063)-exp(2.887)   #comparison between LOS and WBS    On average there's 14 more shrubs in LOS
 plots than WBS plots
```

```
## [1] -13.96173
```

```
exp(2.887-1.5063)-exp(2.887-0.7701) #comparison between WBS and MBS  On average there's 4 more shrubs in MB
S plos than WBS plots
```

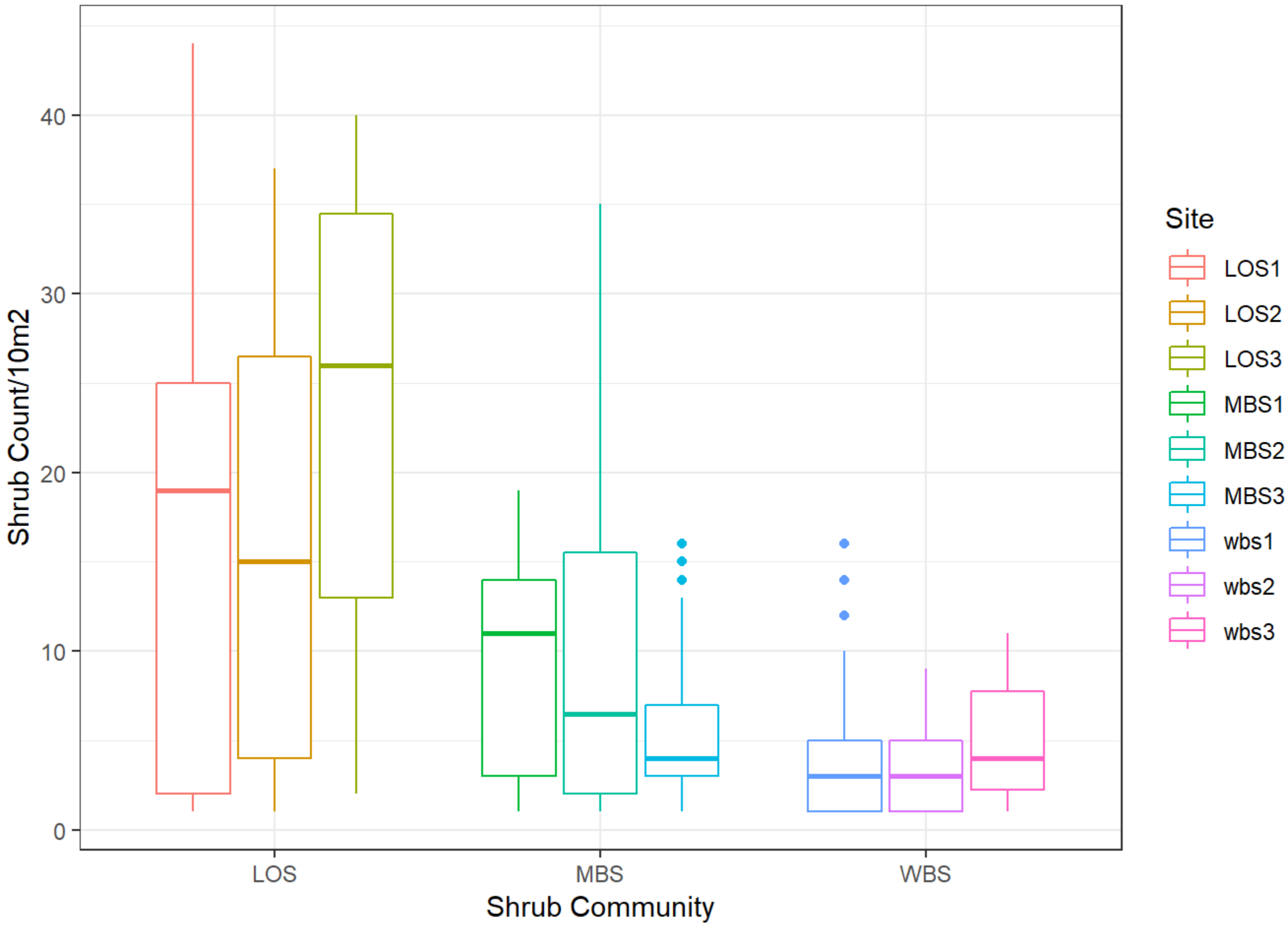
```
## [1] -4.327666
```

```
confint(shrub_mod)  #Site location has a significant effect on the density of shrubs per 10m2
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  2.8384172  2.9354265
## LocationMBS -0.8498115 -0.6909882
## LocationWBS -1.6012983 -1.4123797
```

```
#boxplot(Count~Location, data = shrub, xlab = "Shrub Community", ylab = "Shrub Count per 10m2")
shrub_graph <- ggplot(shrub, aes(x = Location, y = Count, colour = Site)) +
  geom_boxplot(notch = FALSE) +
  labs(x = 'Shrub Community', y = 'Shrub Count/10m2') +
  theme_bw()
shrub_graph
```



2016 shrub biomass

```
bio<-read.csv("shrub_biomass_2016.csv")
head(bio)
```

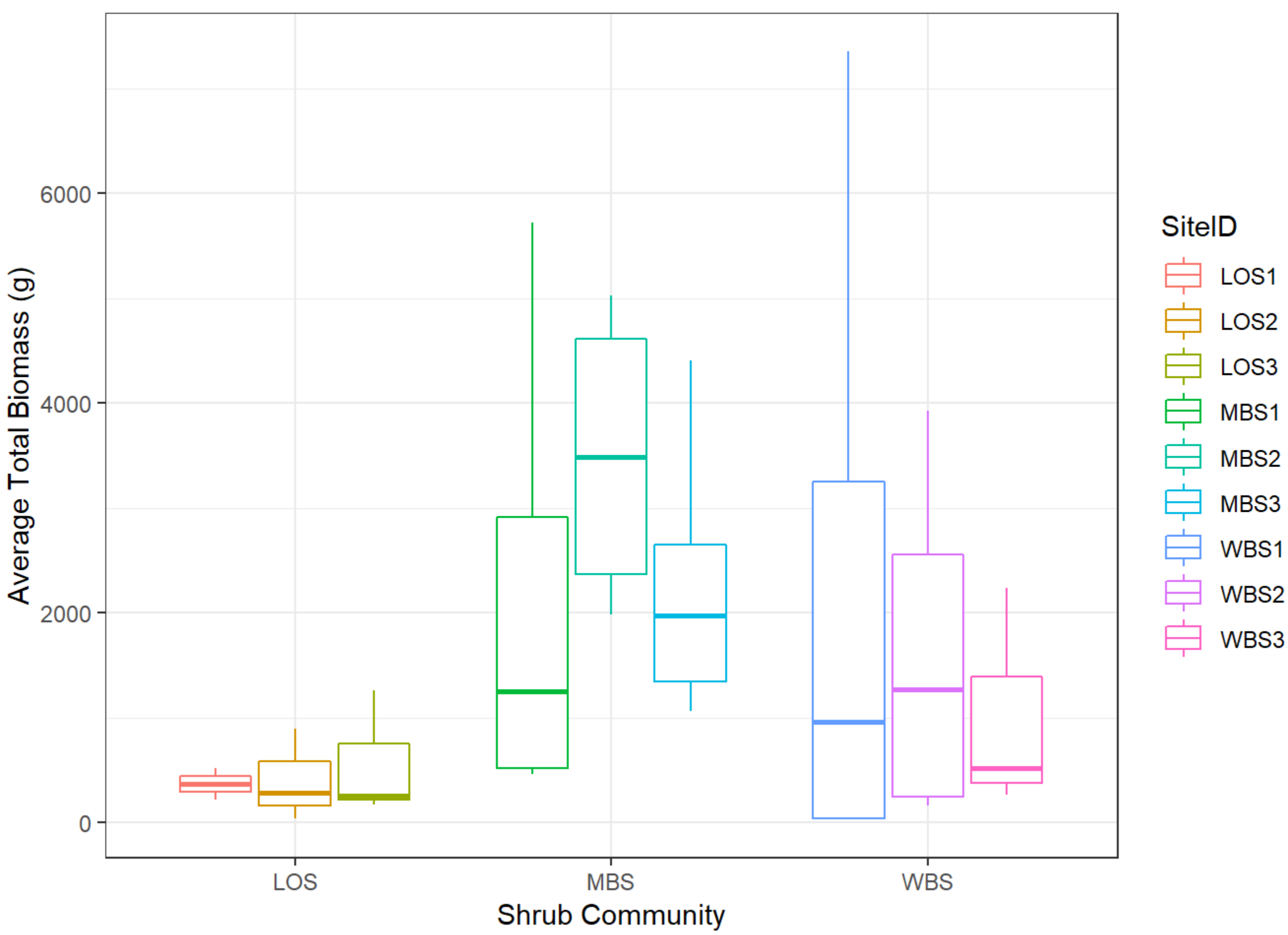
```
##        Date Observer Recorder SiteID SiteName Plot Species SizeClass
## 1 6/21/2016   jordan     cami   LOS1      LSC    6   ARAR8         s
## 2 6/21/2016   jordan     cami   LOS1      LSC    6   ARAR8         m
## 3 6/21/2016   jordan     cami   LOS1      LSC    6   ARAR8         l
## 4 7/18/2016     cami   jordan   LOS2           NA   ARAR8         s
## 5 7/18/2016     cami   jordan   LOS2           NA   ARAR8         m
## 6 7/18/2016     cami   jordan   LOS2           NA   ARAR8         l
##   Height CrownDepth MaxDia MaxPerpDia MinDia CrownDensityClass
## 1     37         30     53         40     34                 6
## 2     40         35     80         70     40                 6
## 3     53         40     80         70     50                 6
## 4     31         19     27         25     20                 6
## 5     34         11     54         47     26                 5
## 6     51         14     84         64     41                 5
##   CrownDensity. ConVxHull WoodBiomass PerGreenBiomass CYGreenBiomass
## 1         91.66     78440      175.38            8.02          34.73
## 2         91.66    224000          NA              NA             NA
## 3         91.66    296800      432.46           23.60          64.90
## 4         91.66     20925      725.56           55.79         113.36
## 5         75.00     86292      232.95           19.78          30.98
## 6         75.00    274176       33.34            5.02           5.31
##   TotalBiomas Location
## 1      218.13      LOS
## 2          NA      LOS
## 3      520.96      LOS
## 4      894.71      LOS
## 5      283.71      LOS
## 6       43.67      LOS
```

```
str(bio)
```

```
## 'data.frame':    45 obs. of  21 variables:
##  $ Date             : Factor w/ 7 levels "5/24/2016","6/13/2016",..: 3 3 3 5 5 5 5 5 5 1 ...
##  $ Observer         : Factor w/ 4 levels "Alex B","alex d",..: 4 4 4 3 3 3 3 3 3 3 ...
##  $ Recorder         : Factor w/ 4 levels "alex b","cami",..: 2 2 2 3 3 3 3 3 3 3 ...
##  $ SiteID           : Factor w/ 9 levels "LOS1","LOS2",..: 1 1 1 2 2 2 3 3 3 7 ...
##  $ SiteName         : Factor w/ 4 levels "","LSC","Nancys",..: 2 2 2 1 1 1 1 1 1 3 ...
##  $ Plot             : int  6 6 6 NA NA NA NA NA NA 2 ...
##  $ Species          : Factor w/ 4 levels "ARAR8","ARTRV",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ SizeClass        : Factor w/ 3 levels "l","m","s": 3 2 1 3 2 1 3 2 1 3 ...
##  $ Height           : int  37 40 53 31 34 51 51 48 70 34 ...
##  $ CrownDepth       : int  30 35 40 19 11 14 30 23 20 31 ...
##  $ MaxDia           : int  53 80 80 27 54 84 38 39 87 34 ...
##  $ MaxPerpDia       : int  40 70 70 25 47 64 33 46 105 20 ...
##  $ MinDia           : int  34 40 50 20 26 41 14 33 25 10 ...
##  $ CrownDensityClass: int  6 6 6 6 5 5 6 6 5 4 ...
##  $ CrownDensity.    : num  91.7 91.7 91.7 91.7 75 ...
##  $ ConVxHull        : int  78440 224000 296800 20925 86292 274176 63954 86112 639450 23120 ...
##  $ WoodBiomass      : num  175 NA 432 726 233 ...
##  $ PerGreenBiomass  : num  8.02 NA 23.6 55.79 19.78 ...
##  $ CYGreenBiomass   : num  34.7 NA 64.9 113.4 31 ...
##  $ TotalBiomas      : num  218 NA 521 895 284 ...
##  $ Location         : Factor w/ 3 levels "LOS","MBS","WBS": 1 1 1 1 1 1 1 1 1 3 ...
```

```
#ggplot_shiny(bio_edit)
#boxplot(TotalBiomas~Location, data = bio, xlab = "Shrub Community", ylab = "Total Biomass (oven dry g)")

#taking out an extreme outlier--> seems like data entry error bc dry shrubs don't weigh 27lbs
bio_edit<-bio[-c(24),]

biomass_graph <- ggplot(bio_edit, aes(x = Location, y = TotalBiomas, colour = SiteID)) +
  geom_boxplot(notch = FALSE) +
  labs(x = 'Shrub Community', y = 'Average Total Biomass (g)') +
  theme_bw()
biomass_graph
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```




```
bio_mod<-glm(TotalBiomas~Location, data = bio_edit, family = Gamma(link = "log"))
coef(bio_mod)
```

```
## (Intercept) LocationMBS LocationWBS
##    6.124306    1.748742    1.212540
```

```
confint(bio_mod)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %   97.5 %
## (Intercept) 5.4937418 6.922420
## LocationMBS 0.8053261 2.627168
## LocationWBS 0.2862579 2.060037
```

```
exp(6.124) #baseline of average shrub dry weight = 456.69 g (LOS)
```

```
## [1] 456.6878
```

```
exp(1.749) #There's a 57% increase in average shrub dry weight between LOS to MBS
```

```
## [1] 5.748851
```

```
        exp(6.124+1.7487)-exp(6.124) #the average dry shrub wgt at MBS = 2167.95 g
```

```
## [1] 2167.955
```

```
exp(1.212) #There's a 34% increase in average shrub dry weight between LOS to WBS
```

```
## [1] 3.360198
```

```
        exp(6.124+1.213)-exp(6.124) #the average shrub dry wgt at WBS = 1079.41 g
```

```
## [1] 1079.409
```

```
#Compare averages of biomass per 10m2 plot--> which sites have the most biomass? Least?
```

```
#wbs<-read.csv("Veg_2018_WBS_NestedOnly.csv")
```