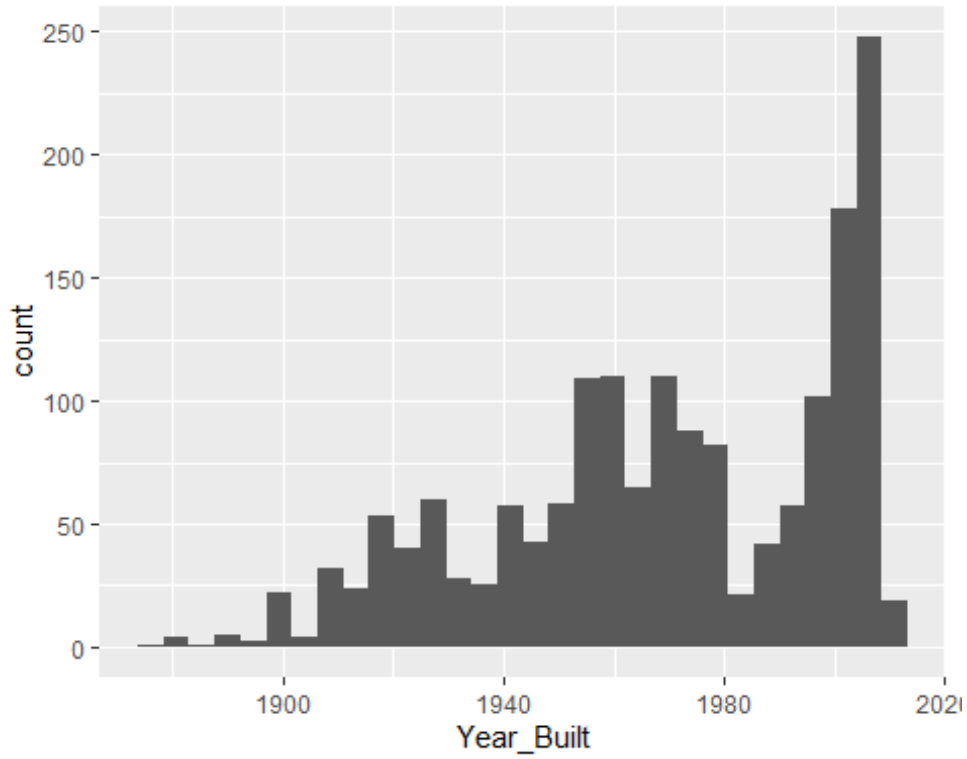
Project Phase 1

Engel, Alec

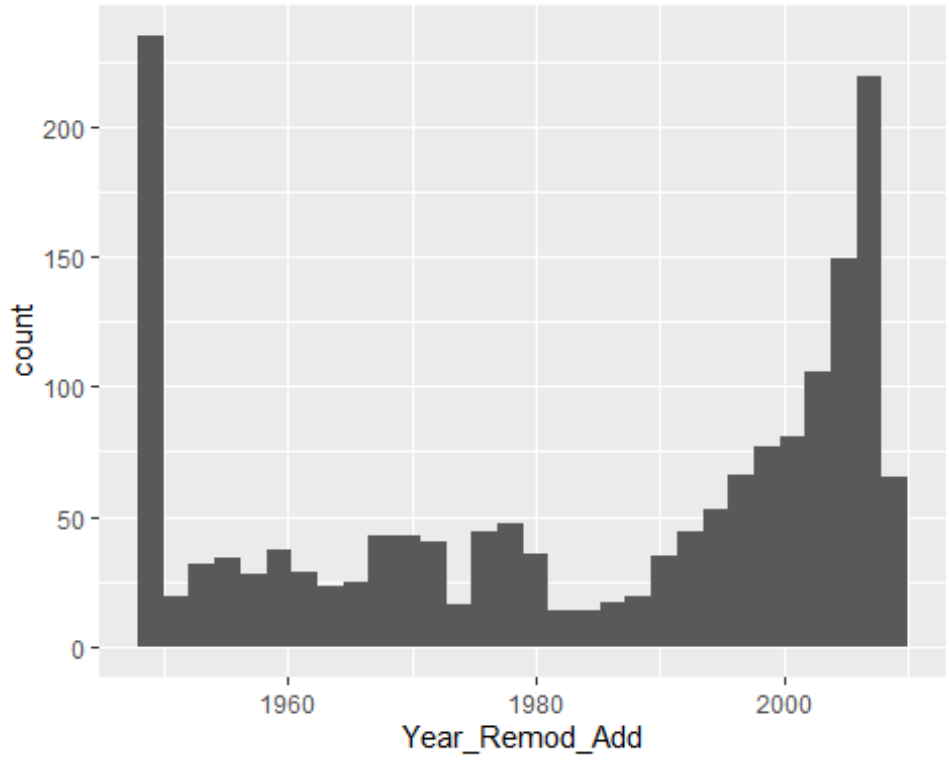
```
# summary(student)
# glimpse(student)
# skim(student)

student = student %>% dplyr::select(-X1) %>%
  mutate_if(is.character,as_factor) %>%
  mutate(Mo_Sold = as_factor(Mo_Sold)) %>%
  mutate(Mo_Sold = fct_recode(Mo_Sold, "Jan" = "1", "Feb" = "2", "Mar" = "3",
"Apr" = "4", "May" = "5", "Jun" = "6",
                              "Jul" = "7", "Aug" = "8", "Sep" = "9", "Oct" =
"10", "Nov" = "11", "Dec" = "12")) %>%
  mutate(BsmtFin_SF_1 = Total_Bsmt_SF - BsmtFin_SF_2 - Bsmt_Unf_SF)

# ggplot(student, aes(Lot_Frontage)) + geom_histogram()
# ggplot(student, aes(Lot_Area)) + geom_histogram()
# ggplot(student, aes(Year_Built)) + geom_histogram()
# ggplot(student, aes(Year_Remod_Add)) + geom_histogram()
# ggplot(student, aes(Mas_Vnr_Area)) + geom_histogram()
# ggplot(student, aes(BsmtFin_SF_1)) + geom_histogram()
# ggplot(student, aes(BsmtFin_SF_2)) + geom_histogram()
# ggplot(student, aes(Bsmt_Unf_SF)) + geom_histogram()
# ggplot(student, aes(Total_Bsmt_SF)) + geom_histogram()
# ggplot(student, aes(First_Flr_SF)) + geom_histogram()
# ggplot(student, aes(Second_Flr_SF)) + geom_histogram()
# ggplot(student, aes(Low_Qual_Fin_SF)) + geom_histogram()
# ggplot(student, aes(Gr_Liv_Area)) + geom_histogram()
# ggplot(student, aes(Bsmt_Full_Bath)) + geom_histogram()
# ggplot(student, aes(Bsmt_Half_Bath)) + geom_histogram()
# ggplot(student, aes(Full_Bath)) + geom_histogram()
# ggplot(student, aes(Half_Bath)) + geom_histogram()
# ggplot(student, aes(Bedroom_AbvGr)) + geom_histogram()
# ggplot(student, aes(Kitchen_AbvGr)) + geom_histogram()
# ggplot(student, aes(TotRms_AbvGrd)) + geom_histogram()
# ggplot(student, aes(Fireplaces)) + geom_histogram()
# ggplot(student, aes(Kitchen_AbvGr)) + geom_histogram()
# ggplot(student, aes(TotRms_AbvGrd)) + geom_histogram()
# ggplot(student, aes(Fireplaces)) + geom_histogram()
# ggplot(student, aes(Garage_Cars)) + geom_histogram()
# ggplot(student, aes(Garage_Area)) + geom_histogram()
# ggplot(student, aes(Wood_Deck_SF)) + geom_histogram()
# ggplot(student, aes(Open_Porch_SF)) + geom_histogram()
# ggplot(student, aes(Enclosed_Porch)) + geom_histogram()
# ggplot(student, aes(Three_season_porch)) + geom_histogram()
# ggplot(student, aes(Screen_Porch)) + geom_histogram()
```

```r
# ggplot(student, aes(Pool_Area)) + geom_histogram()
# ggplot(student, aes(Misc_Val)) + geom_histogram()
# ggplot(student, aes(Year_Sold)) + geom_histogram()
# ggplot(student, aes(Longitude)) + geom_histogram()
# ggplot(student, aes(Latitude)) + geom_histogram()

student = student %>% filter(Lot_Frontage < 175) %>%
  filter(Lot_Area < 40000) %>%
  filter(Mas_Vnr_Area < 400) %>%
  filter(BsmtFin_SF_1 < 1600) %>%
  filter(BsmtFin_SF_2 < 400) %>%
  filter(Bsmt_Unf_SF < 2250) %>%
  filter(Total_Bsmt_SF < 2750) %>%
  filter(Bsmt_Half_Bath < 1.1) %>%
  filter(Full_Bath > 0) %>%
  filter(Half_Bath < 1.1) %>%
  filter(Pool_Area < 1) %>%
  filter(First_Flr_SF < 2750) %>%
  filter(Second_Flr_SF < 1400) %>%
  filter(Low_Qual_Fin_SF < 600) %>%
  filter(Gr_Liv_Area < 3750) %>%
  filter(Kitchen_AbvGr < 3) %>%
  filter(Fireplaces < 4) %>%
  filter(Garage_Cars < 4) %>%
  filter(Garage_Area < 1250) %>%
  filter(Wood_Deck_SF < 550) %>%
  filter(Open_Porch_SF < 350) %>%
  filter(Enclosed_Porch < 300) %>%
  filter(Three_season_porch < 240) %>%
  filter(Screen_Porch < 400) %>%
  filter(Misc_Val < 1000)

ggplot(student, aes(Year_Built)) + geom_histogram()

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



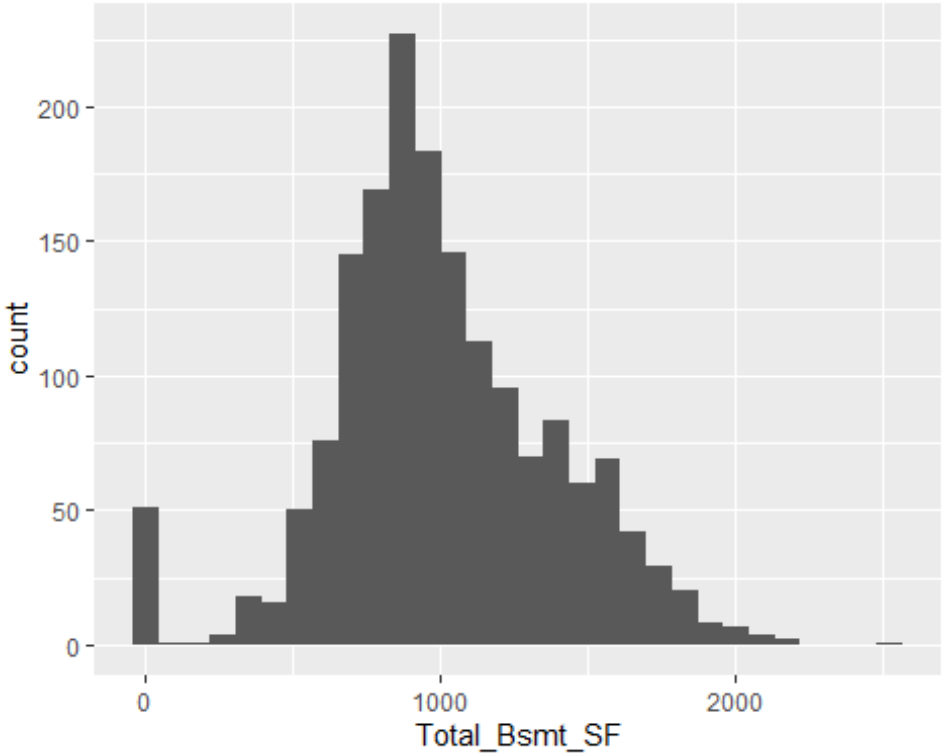

```
ggplot(student, aes(Year_Remod_Add)) + geom_histogram()
```

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



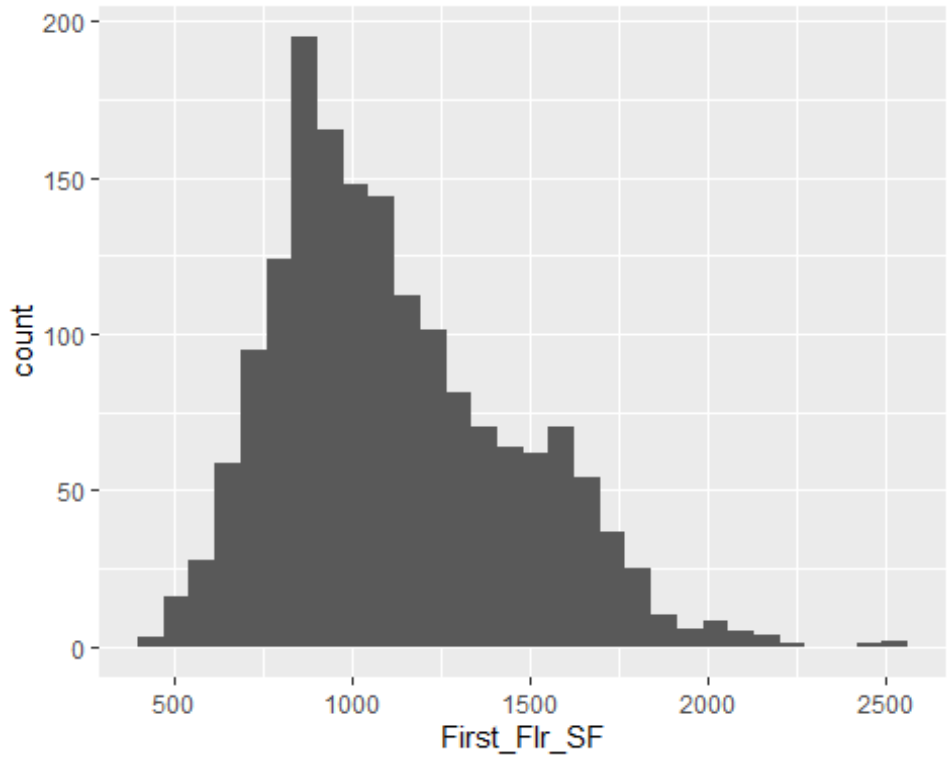
```
ggplot(student, aes(Total_Bsmt_SF)) + geom_histogram()
```

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



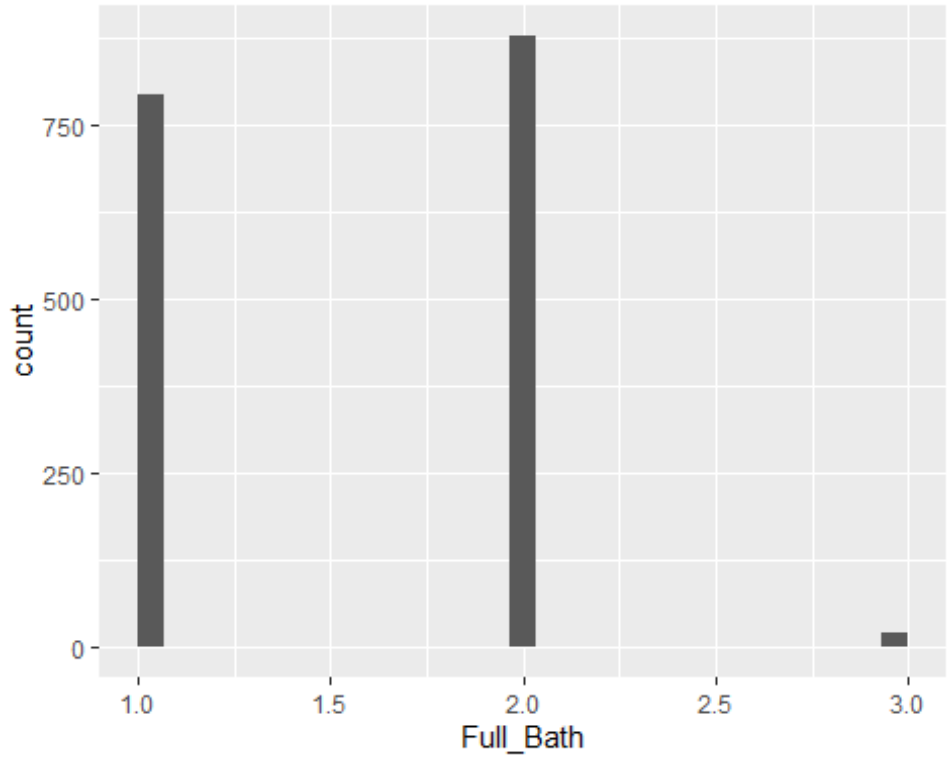

```
ggplot(student, aes(First_Flr_SF)) + geom_histogram()
```

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

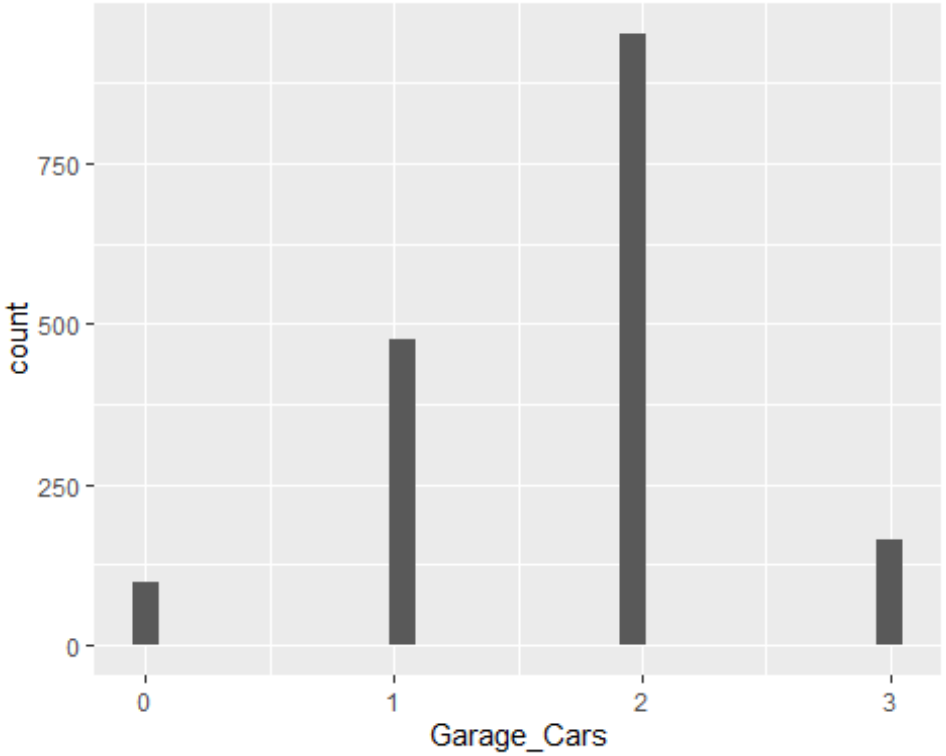



```
ggplot(student, aes(Full_Bath)) + geom_histogram()
```

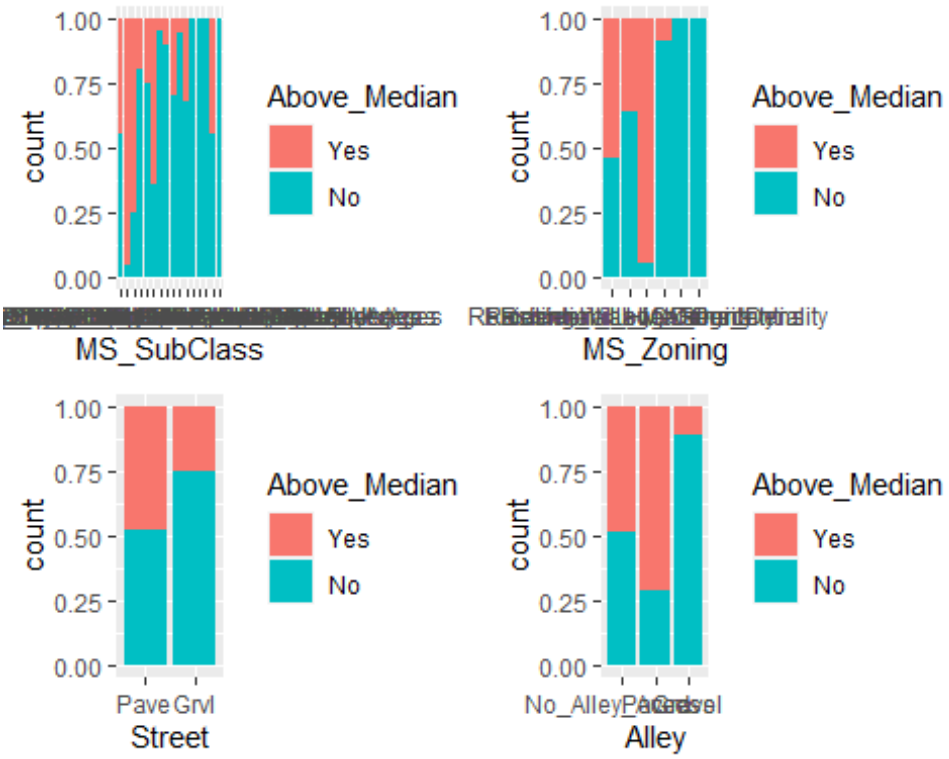
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



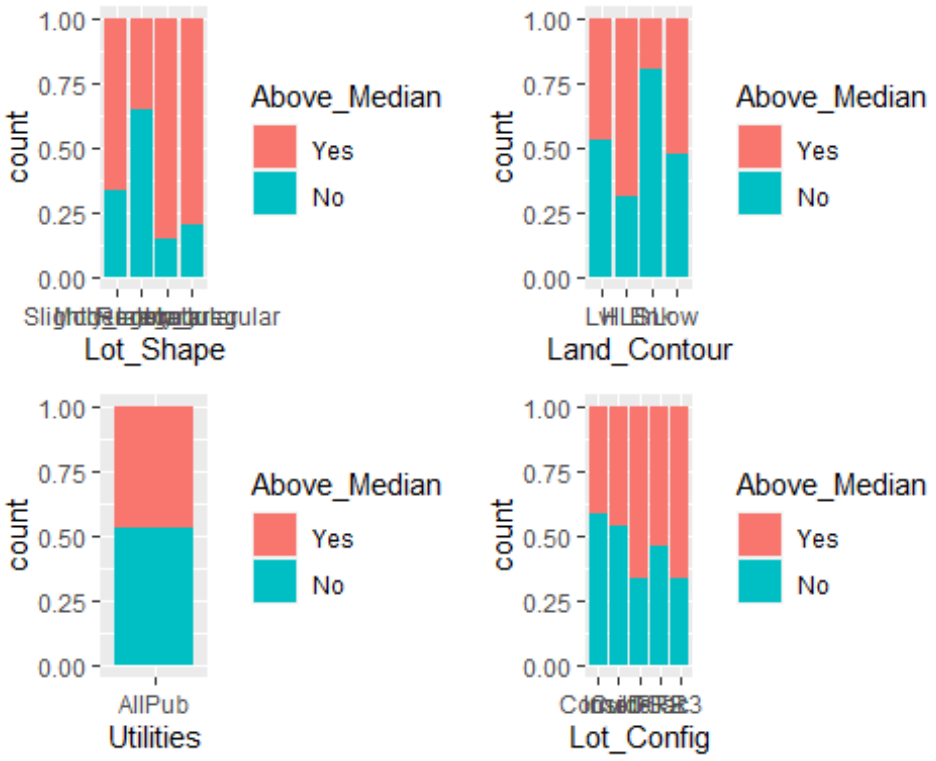
```
ggplot(student, aes(Garage_Cars)) + geom_histogram()

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



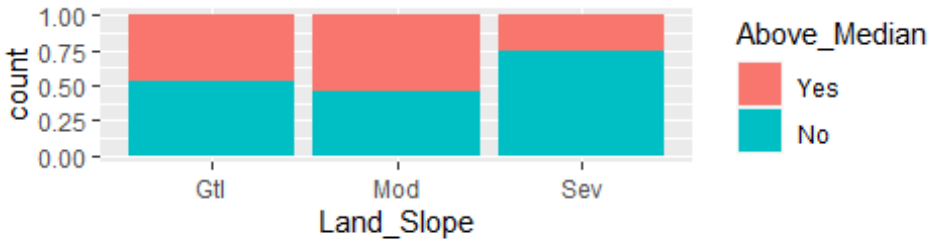

```
set.seed(123)
student_split = initial_split(student, prob = 0.80, strata = Above_Median)
train = training(student_split)
test = testing(student_split)

p1 = ggplot(train, aes(x = MS_SubClass, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = MS_Zoning, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Street, fill = Above_Median)) + geom_bar(position
= "fill")
p4 = ggplot(train, aes(x = Alley, fill = Above_Median)) + geom_bar(position =
"fill")
grid.arrange(p1,p2,p3,p4)
```

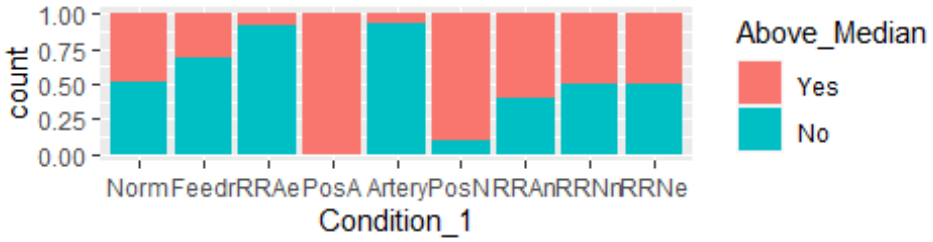


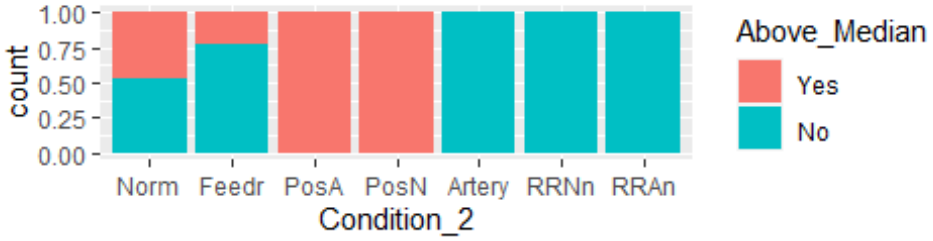

```
p1 = ggplot(train, aes(x = Lot_Shape, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Land_Contour, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Utilities, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Lot_Config, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



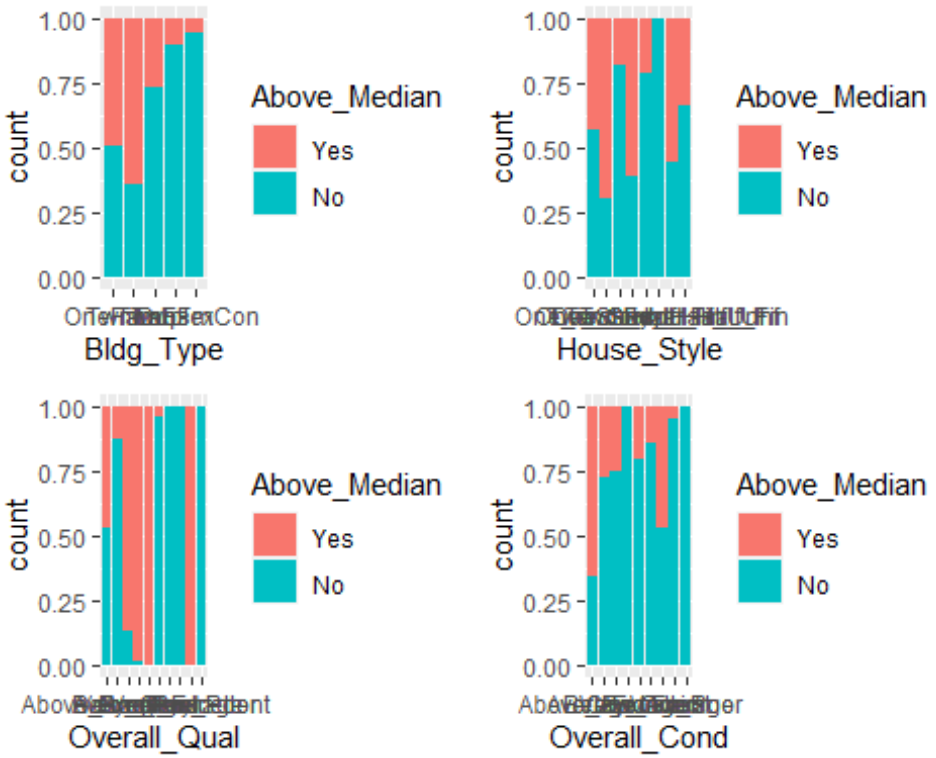

```
p1 = ggplot(train, aes(x = Land_Slope, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Condition_1, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Condition_2, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3)
```



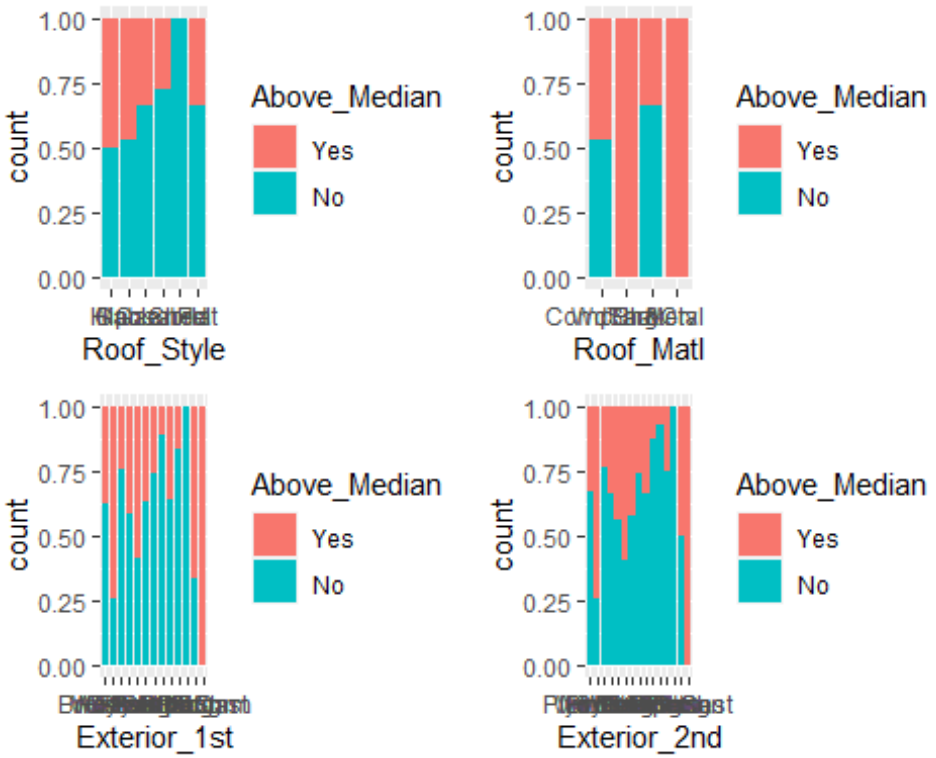




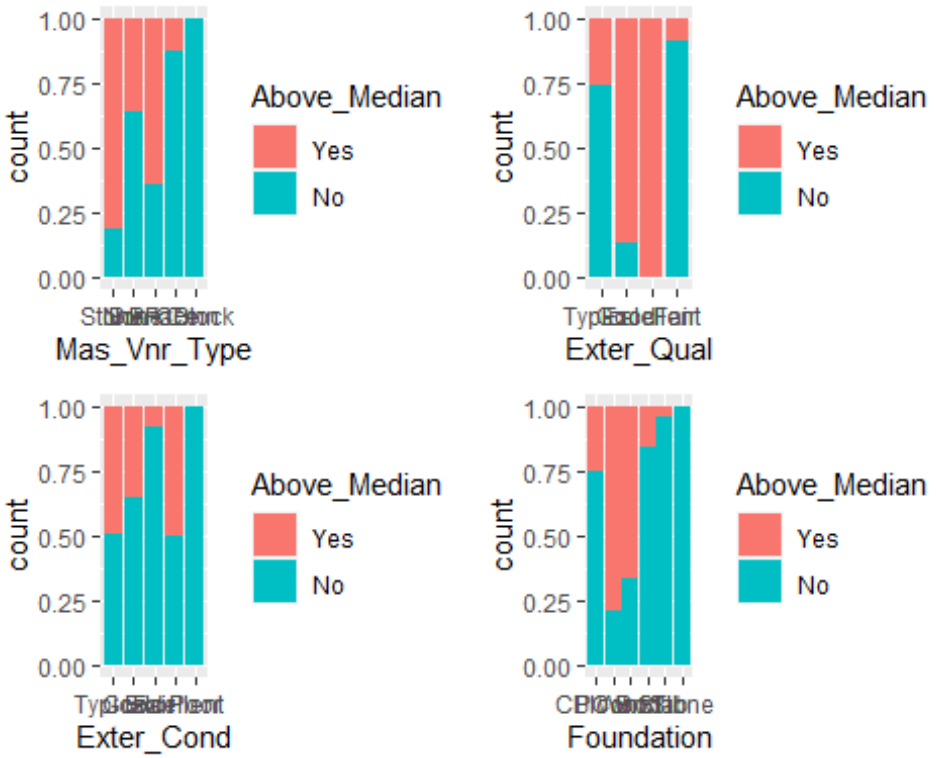

```
p1 = ggplot(train, aes(x = Bldg_Type, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = House_Style, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Overall_Qual, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Overall_Cond, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



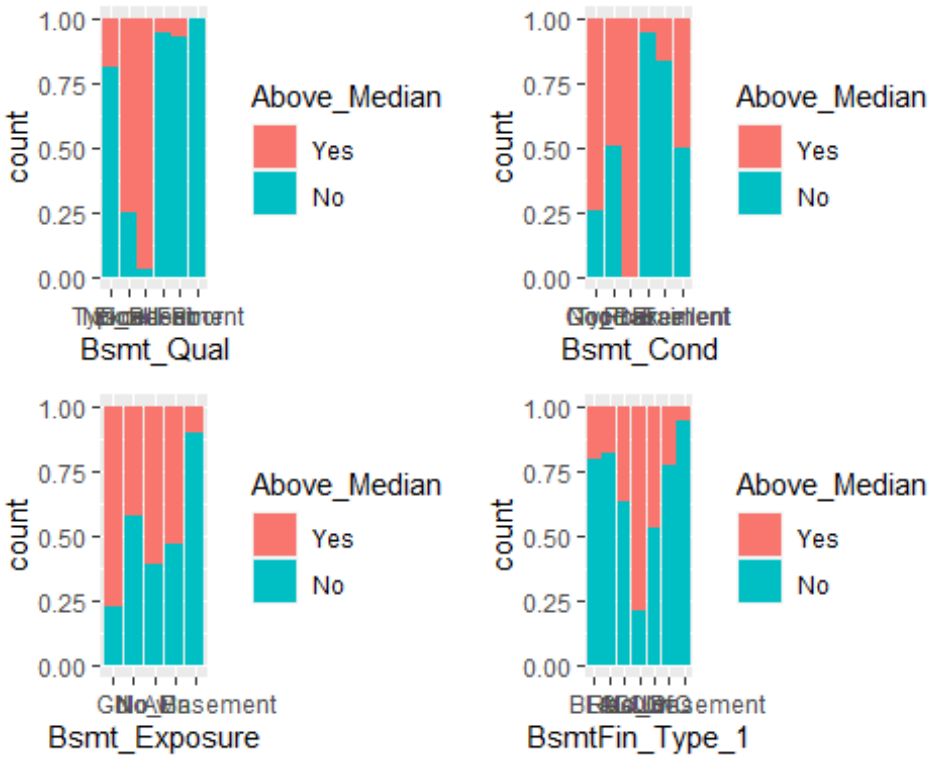

```
p1 = ggplot(train, aes(x = Roof_Style, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Roof_Matl, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Exterior_1st, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Exterior_2nd, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



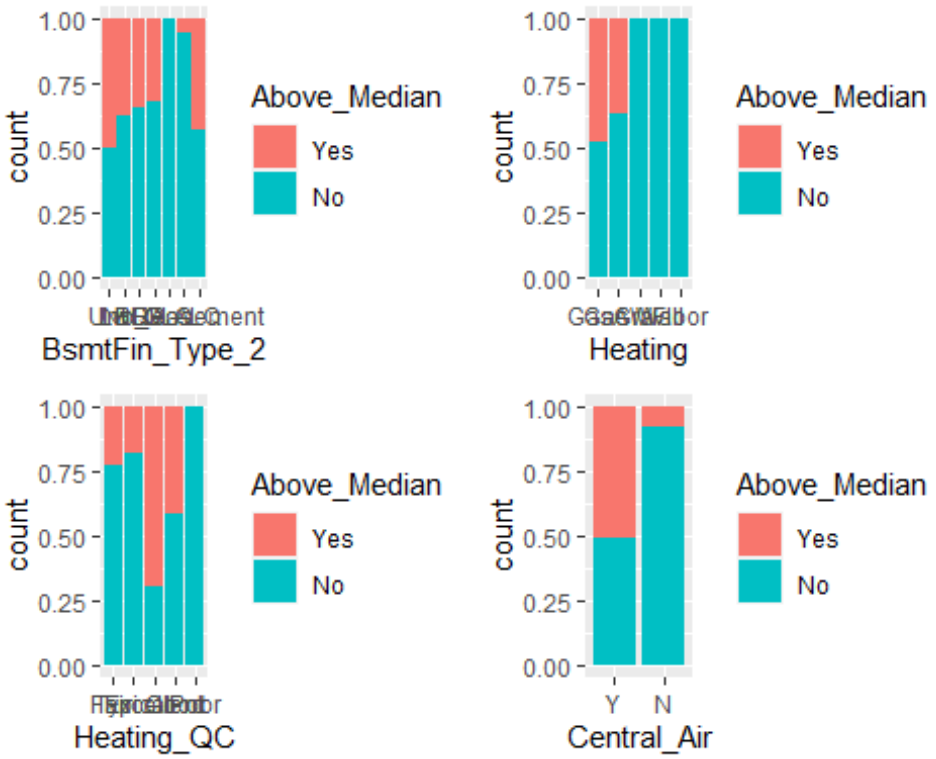

```
p1 = ggplot(train, aes(x = Mas_Vnr_Type, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Exter_Qual, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Exter_Cond, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Foundation, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



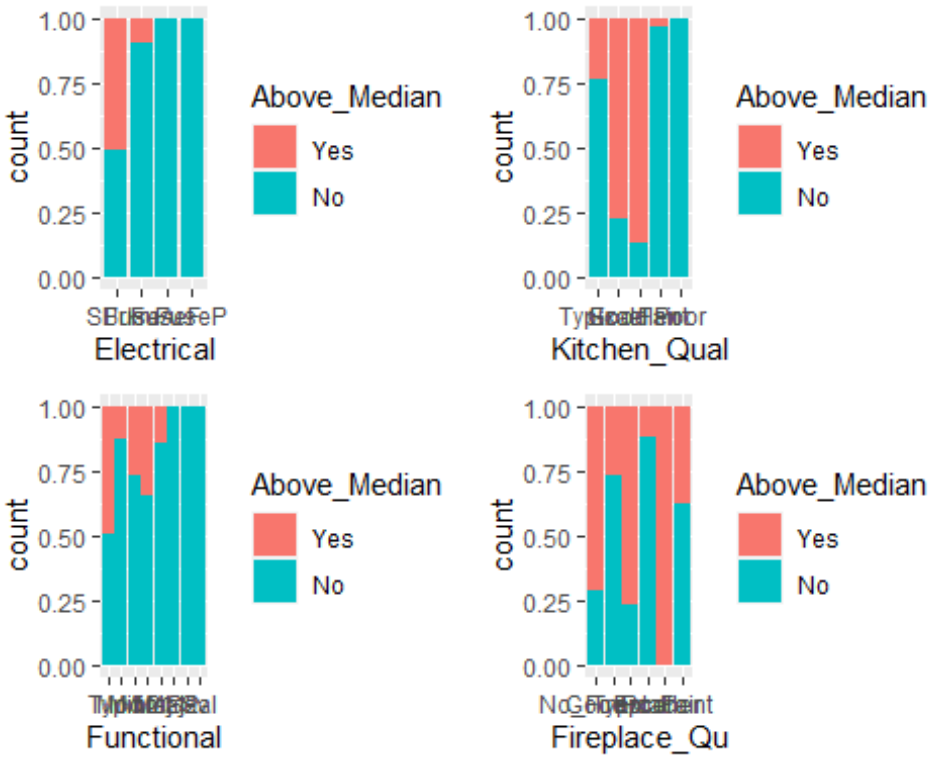

```
p1 = ggplot(train, aes(x = Bsmt_Qual, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Bsmt_Cond, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Bsmt_Exposure, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = BsmtFin_Type_1, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



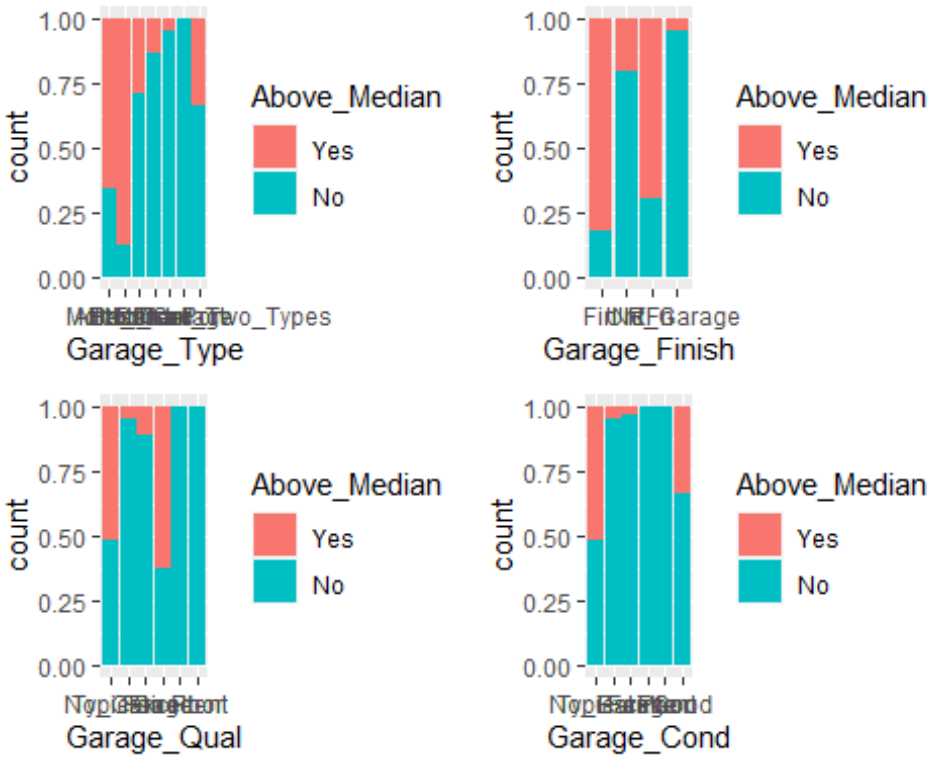

```
p1 = ggplot(train, aes(x = BsmtFin_Type_2, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Heating, fill = Above_Median)) + geom_bar(position
= "fill")
p3 = ggplot(train, aes(x = Heating_QC, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Central_Air, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



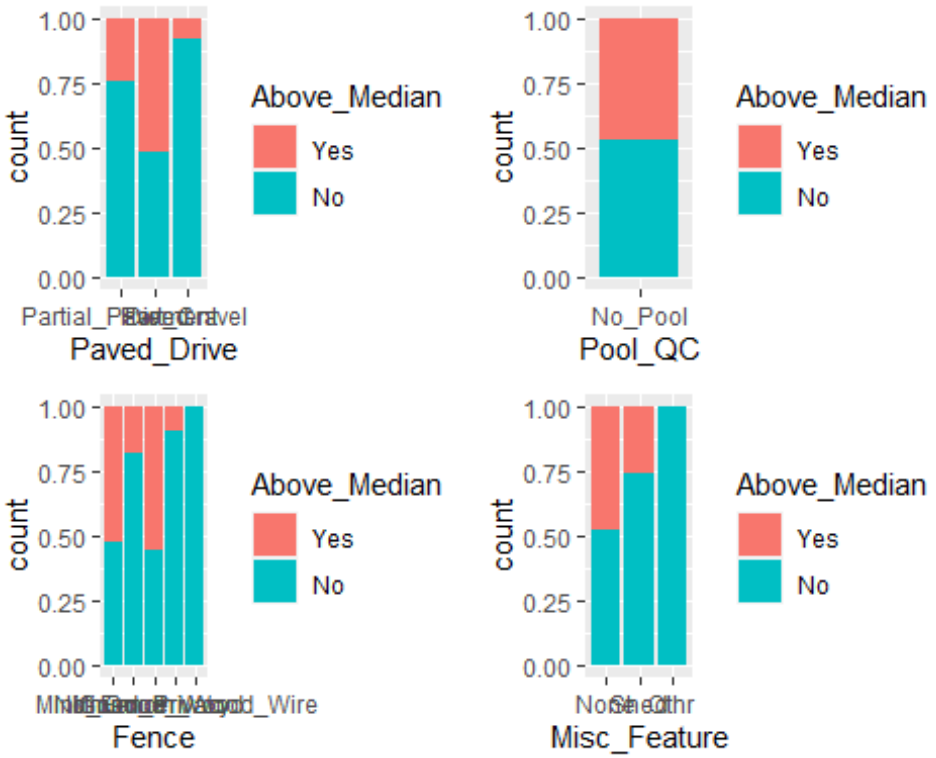

```
p1 = ggplot(train, aes(x = Electrical, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Kitchen_Qual, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Functional, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Fireplace_Qu, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



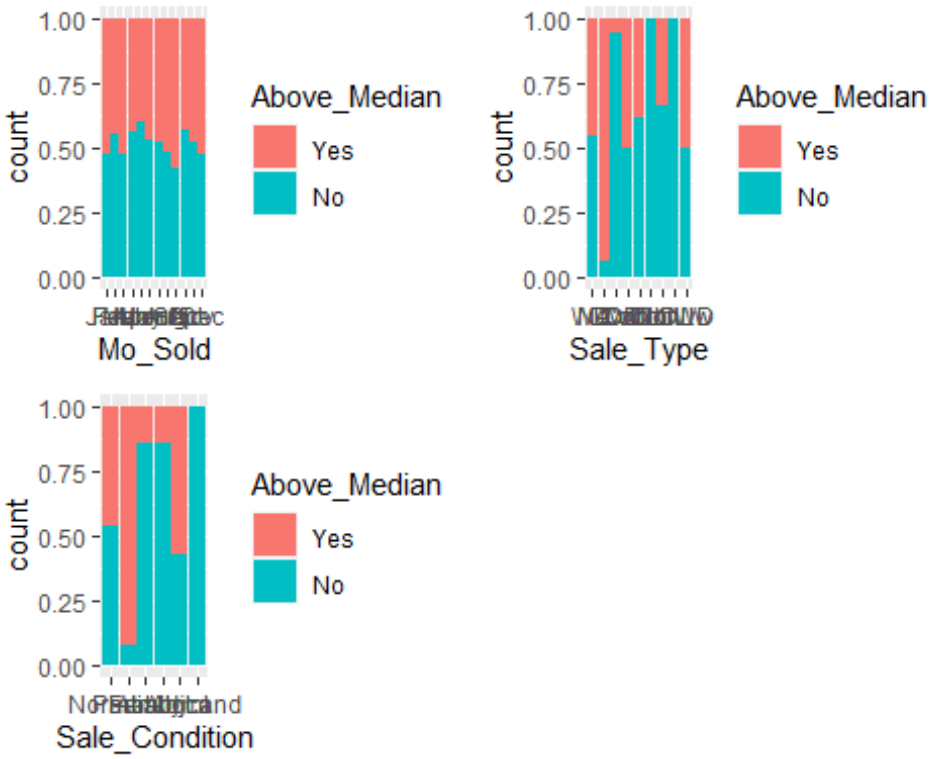

```
p1 = ggplot(train, aes(x = Garage_Type, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Garage_Finish, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Garage_Qual, fill = Above_Median)) +
geom_bar(position = "fill")
p4 = ggplot(train, aes(x = Garage_Cond, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



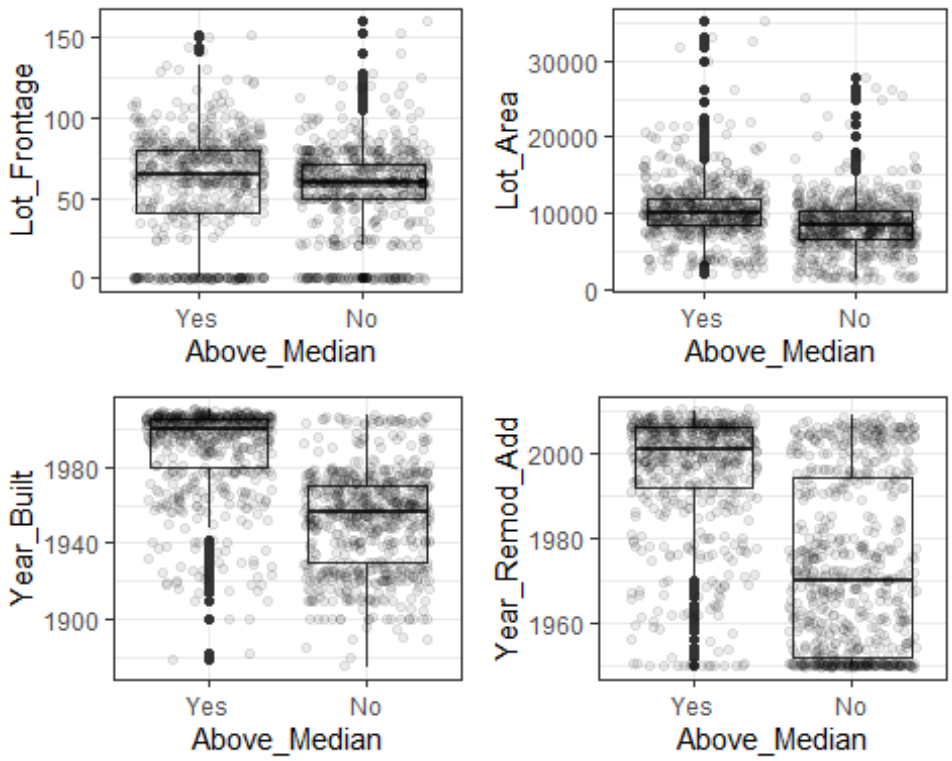

```
p1 = ggplot(train, aes(x = Paved_Drive, fill = Above_Median)) +
geom_bar(position = "fill")
p2 = ggplot(train, aes(x = Pool_QC, fill = Above_Median)) + geom_bar(position
= "fill")
p3 = ggplot(train, aes(x = Fence, fill = Above_Median)) + geom_bar(position =
"fill")
p4 = ggplot(train, aes(x = Misc_Feature, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3,p4)
```



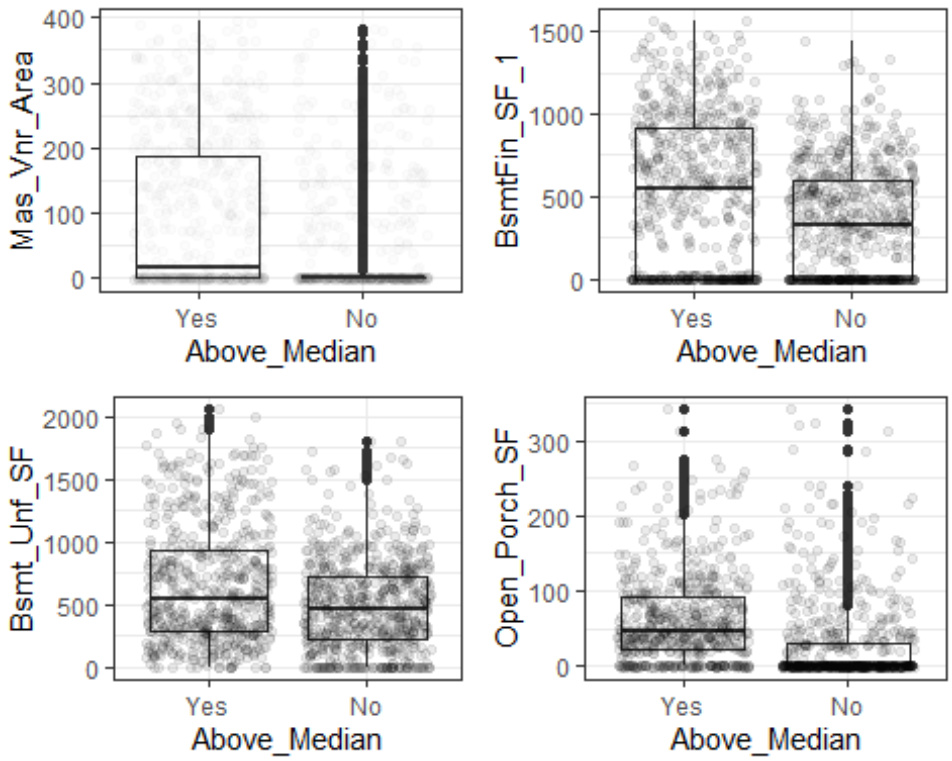

```
p1 = ggplot(train, aes(x = Mo_Sold, fill = Above_Median)) + geom_bar(position
= "fill")
p2 = ggplot(train, aes(x = Sale_Type, fill = Above_Median)) +
geom_bar(position = "fill")
p3 = ggplot(train, aes(x = Sale_Condition, fill = Above_Median)) +
geom_bar(position = "fill")
grid.arrange(p1,p2,p3, ncol = 2)
```



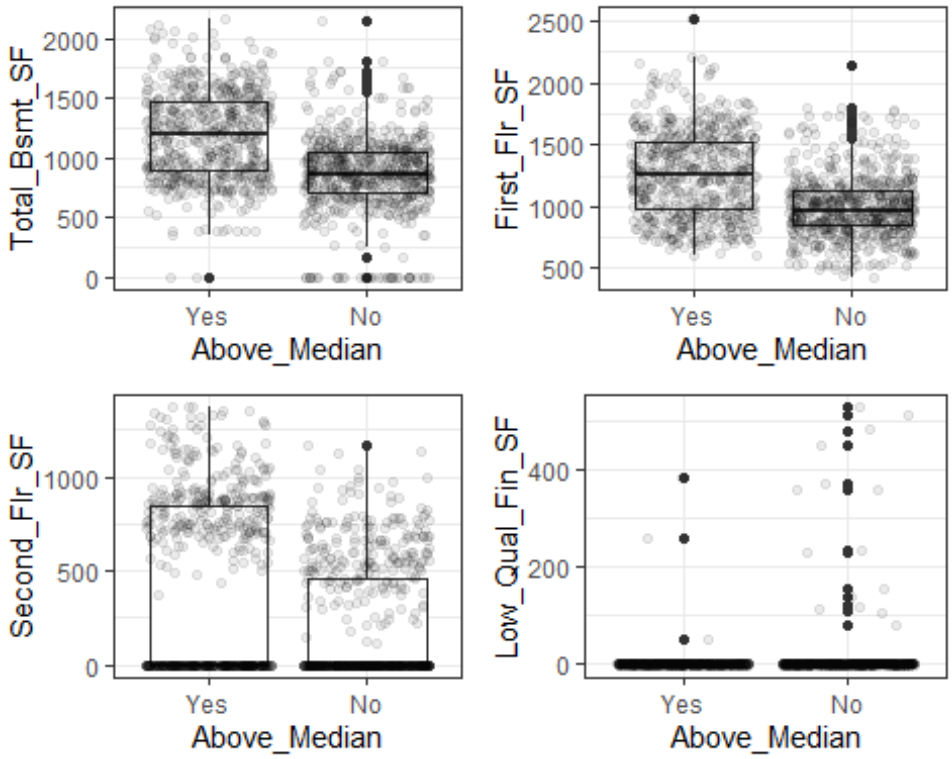

```
p1 = ggplot(train, aes(x = Above_Median, y = Lot_Frontage)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = Lot_Area)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Year_Built)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = Year_Remod_Add)) +
geom_boxplot() + geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```



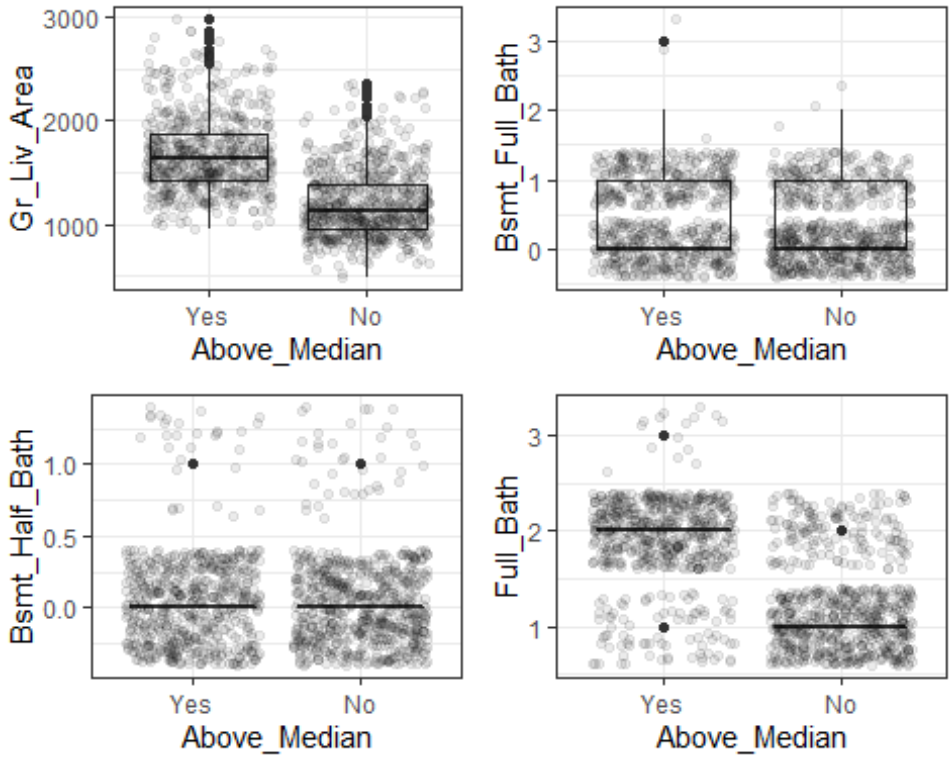

```
p1 = ggplot(train, aes(x = Above_Median, y = Mas_Vnr_Area)) + geom_boxplot()
+ geom_jitter(alpha = 0.01) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = BsmtFin_SF_1)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Bsmt_Unf_SF)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = Open_Porch_SF)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```



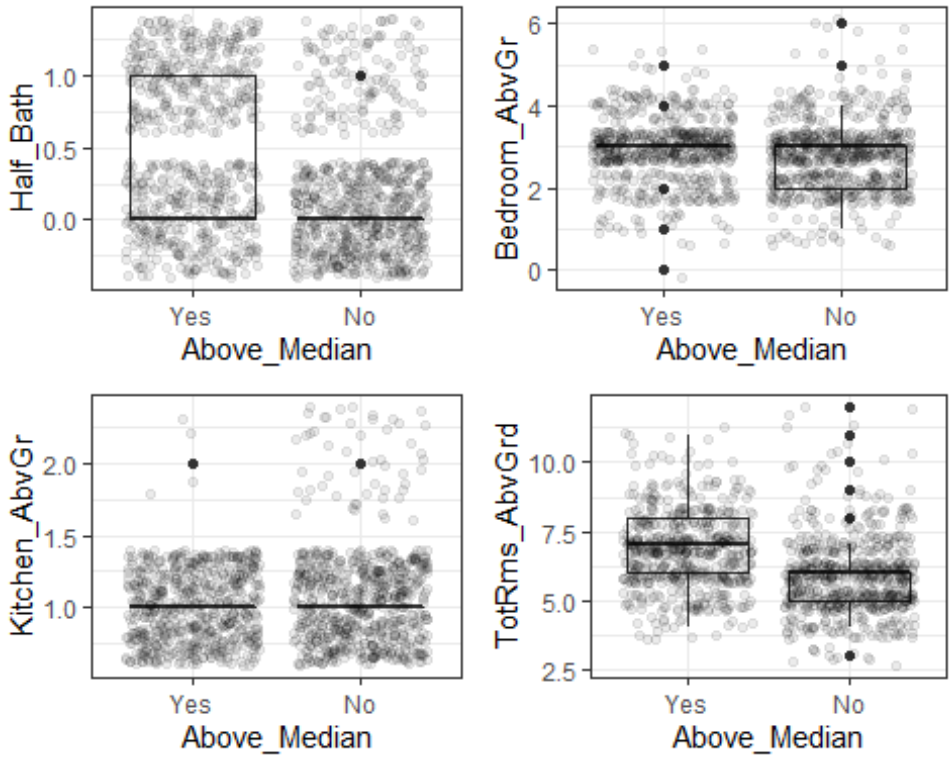

```
p1 = ggplot(train, aes(x = Above_Median, y = Total_Bsmt_SF)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = First_Flr_SF)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Second_Flr_SF)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = Low_Qual_Fin_SF)) +
geom_boxplot() + geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```



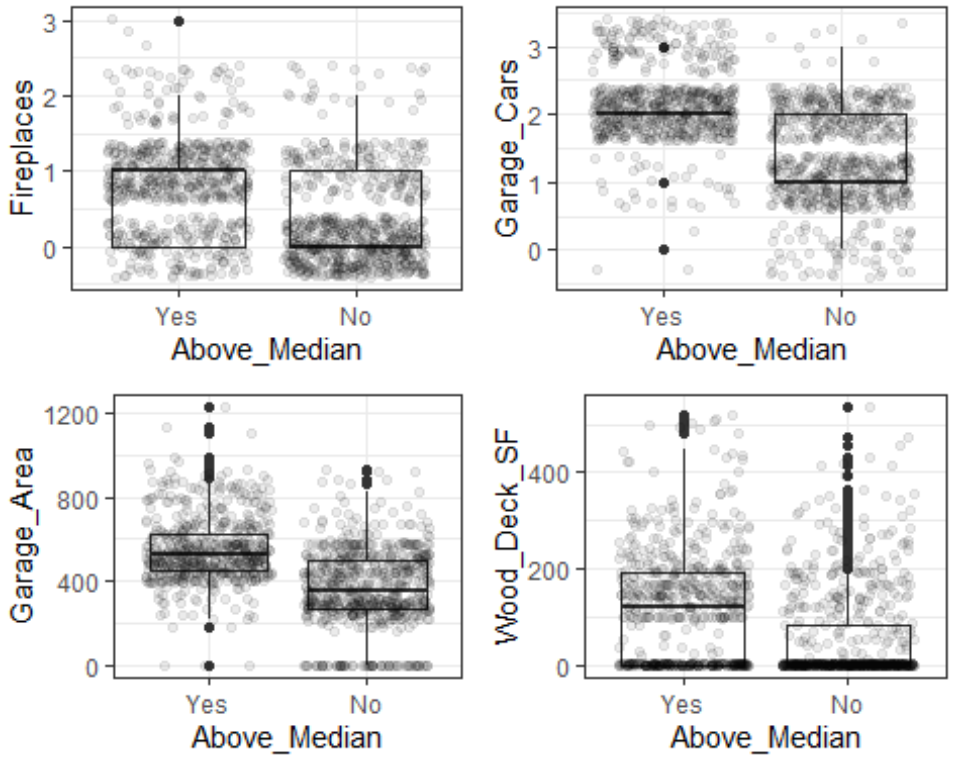

```
p1 = ggplot(train, aes(x = Above_Median, y = Gr_Liv_Area)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = Bsmt_Full_Bath)) +
geom_boxplot() + geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Bsmt_Half_Bath)) +
geom_boxplot() + geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = Full_Bath)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```



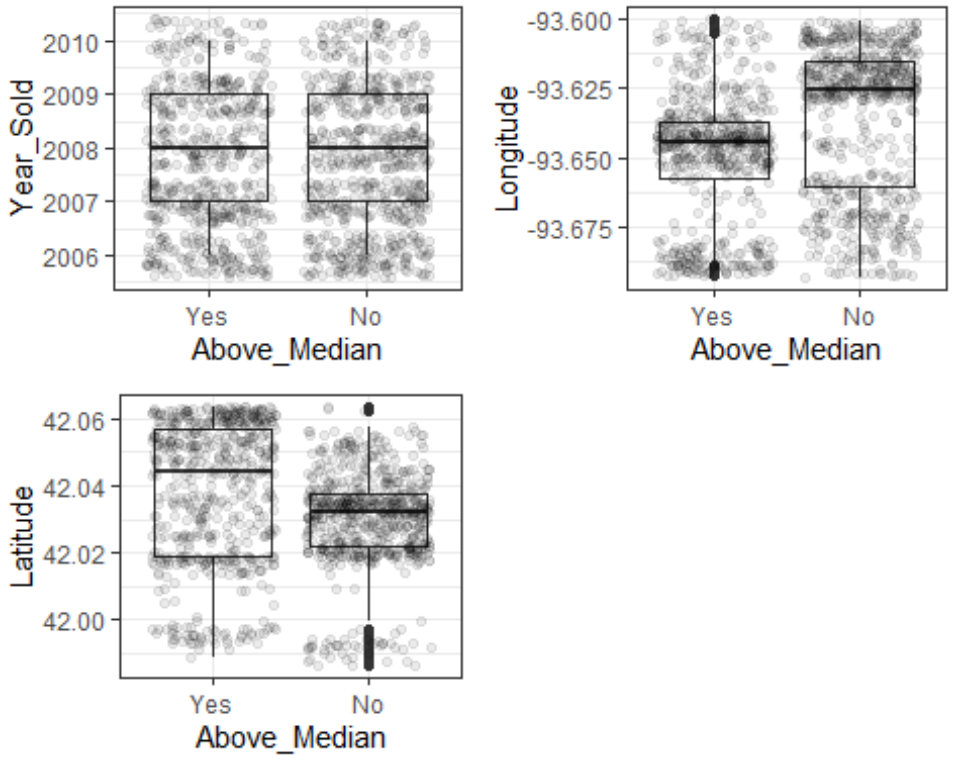

```
p1 = ggplot(train, aes(x = Above_Median, y = Half_Bath)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = Bedroom_AbvGr)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Kitchen_AbvGr)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = TotRms_AbvGrd)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```



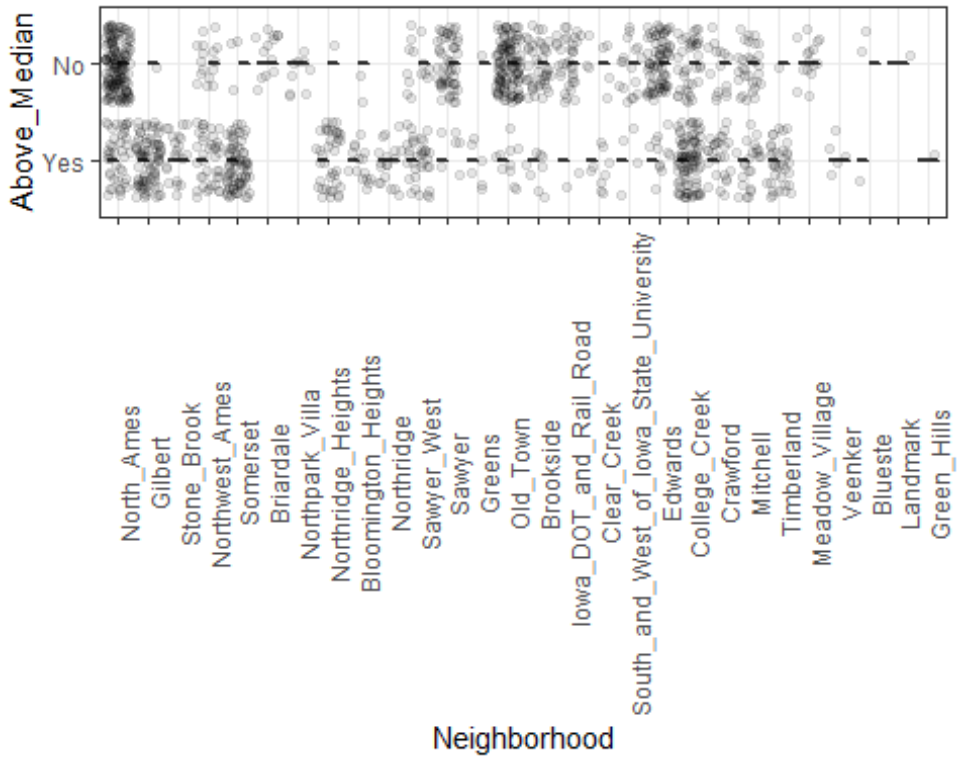

```
p1 = ggplot(train, aes(x = Above_Median, y = Fireplaces)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = Garage_Cars)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Garage_Area)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(train, aes(x = Above_Median, y = Wood_Deck_SF)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```




```
p1 = ggplot(train, aes(x = Above_Median, y = Year_Sold)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(train, aes(x = Above_Median, y = Longitude)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(train, aes(x = Above_Median, y = Latitude)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,ncol = 2)
```




```
ggplot(train, aes(x = Neighborhood, y = Above_Median)) + geom_boxplot() +
geom_jitter(alpha = 0.1) + theme_bw() +
  theme(axis.text.x = element_text(angle = 90))
```



```r
student_recipe = recipe(Above_Median ~., train) %>%
  step_other(Neighborhood,threshold = .02) %>%
  step_other(MS_SubClass,threshold = .02) %>%
  step_other(Overall_Qual,threshold = .02) %>%
  step_other(Overall_Cond,threshold = .02) %>%
  step_other(Exterior_1st,threshold = .02) %>%
  step_other(Exterior_2nd,threshold = .02) %>%
  step_other(Condition_1,threshold = .02) %>%
  step_other(Condition_2,threshold = .02) %>%
  step_other(Functional,threshold = .02) %>%
  step_other(Sale_Type,threshold = .02) %>%
  step_dummy(all_nominal(), -all_outcomes())

rf_model = rand_forest() %>%
  set_engine("ranger", importance = "permutation") %>%
  set_mode("classification")

student_wflow =
  workflow() %>%
  add_model(rf_model) %>%
  add_recipe(student_recipe)

set.seed(123)
student_fit = fit(student_wflow, train)

trainpredrf = predict(student_fit, train)
head(trainpredrf)
```

```
## # A tibble: 6 x 1
##   .pred_class
##   <fct>
## 1 Yes
## 2 No
## 3 Yes
## 4 Yes
## 5 Yes
## 6 Yes
```

```r
confusionMatrix(trainpredrf$.pred_class, train$Above_Median,
                positive = "Yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes 593   7
##        No    7 662
##
##                Accuracy : 0.989
##                  95% CI : (0.9816, 0.994)
```

```
##     No Information Rate : 0.5272
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9779
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9883
##             Specificity : 0.9895
##          Pos Pred Value : 0.9883
##          Neg Pred Value : 0.9895
##              Prevalence : 0.4728
##          Detection Rate : 0.4673
##    Detection Prevalence : 0.4728
##       Balanced Accuracy : 0.9889
##
##        'Positive' Class : Yes
##
```

```
testpredrf = predict(student_fit, test)
head(testpredrf)
```

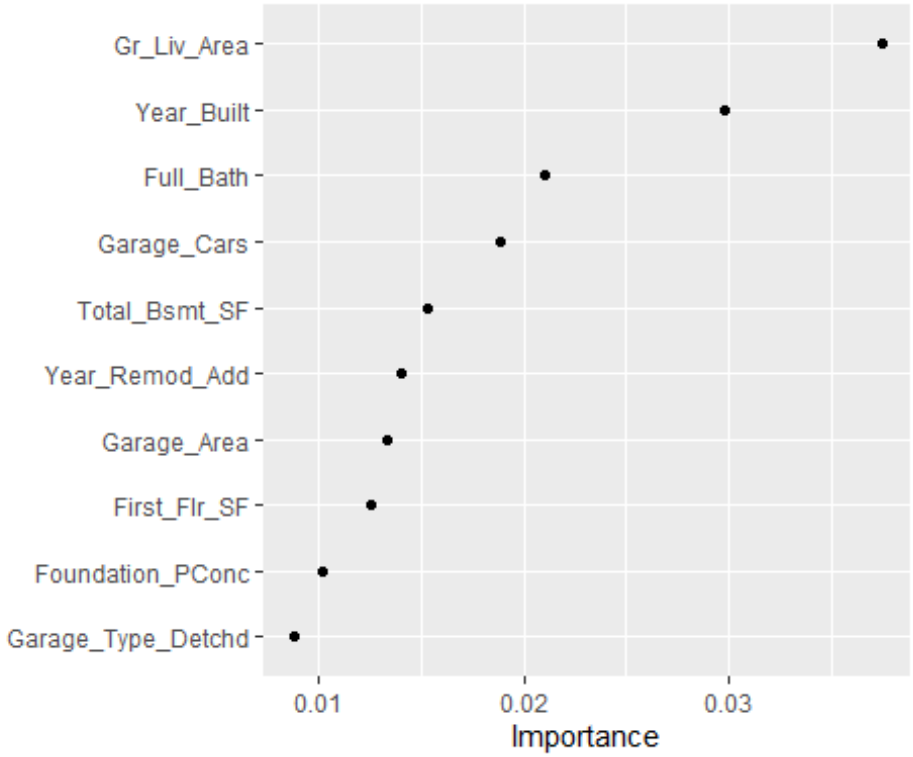
```
## # A tibble: 6 x 1
##   .pred_class
##   <fct>
## 1 Yes
## 2 Yes
## 3 Yes
## 4 No
## 5 Yes
## 6 Yes
```

```
confusionMatrix(testpredrf$.pred_class, test$Above_Median,
                positive = "Yes")
```

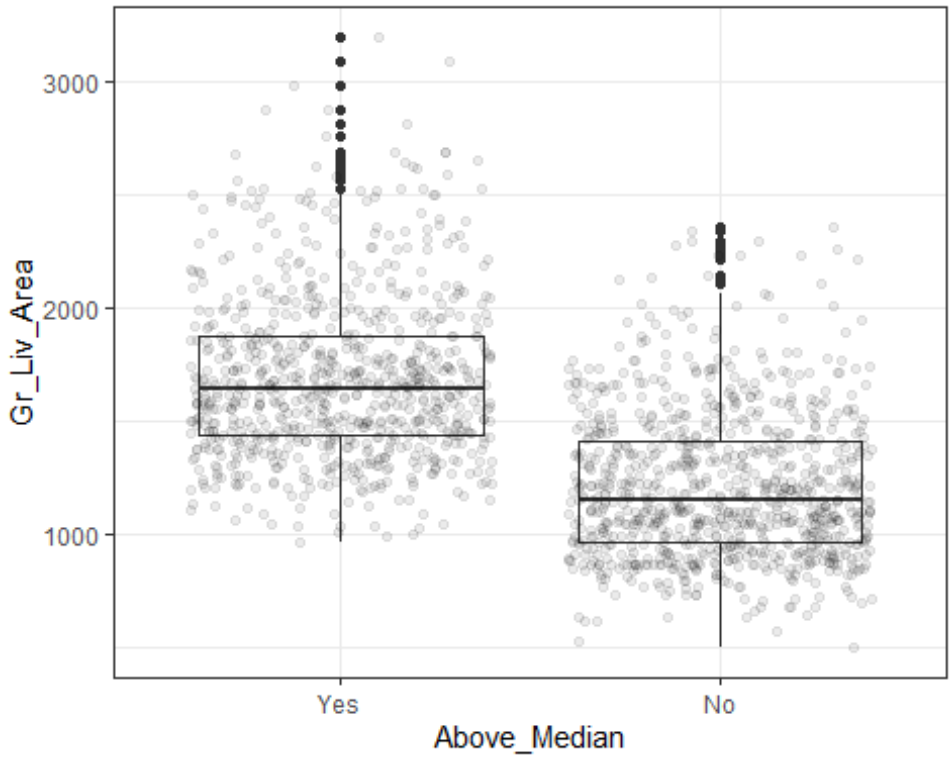
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes 175   6
##        No   24 216
##
##                Accuracy : 0.9287
##                  95% CI : (0.8998, 0.9514)
##     No Information Rate : 0.5273
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8564
##
##  Mcnemar's Test P-Value : 0.001911
##
```

```
##             Sensitivity : 0.8794
##             Specificity : 0.9730
##          Pos Pred Value : 0.9669
##          Neg Pred Value : 0.9000
##              Prevalence : 0.4727
##          Detection Rate : 0.4157
##    Detection Prevalence : 0.4299
##       Balanced Accuracy : 0.9262
##
##        'Positive' Class : Yes
##
```

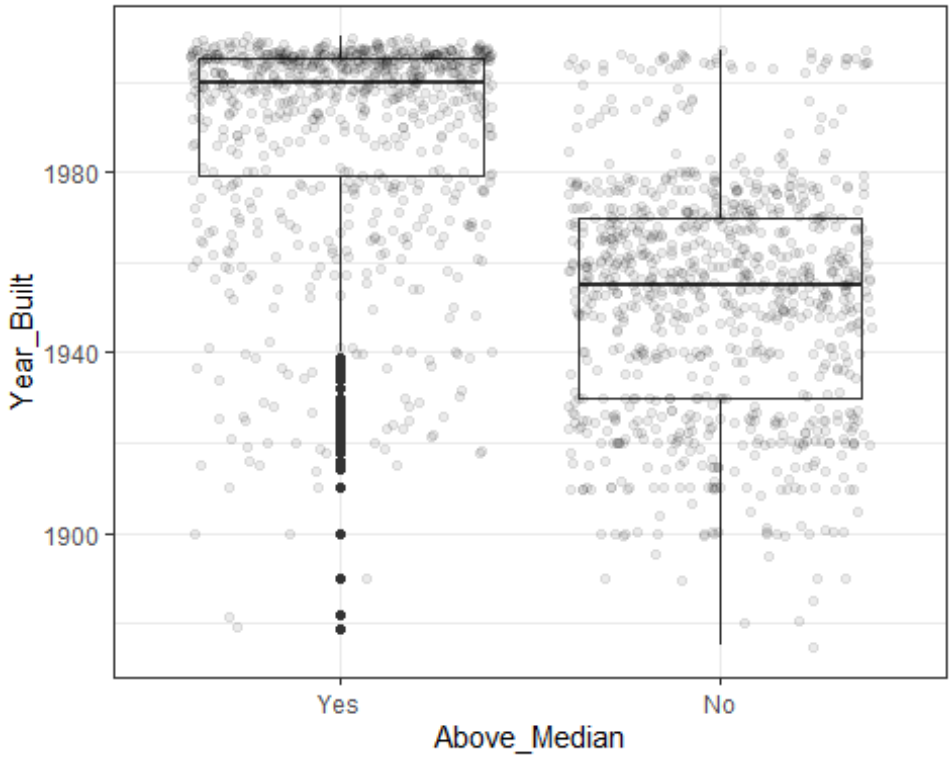
```
student_fit %>% pull_workflow_fit() %>% vip(geom = "point")
```



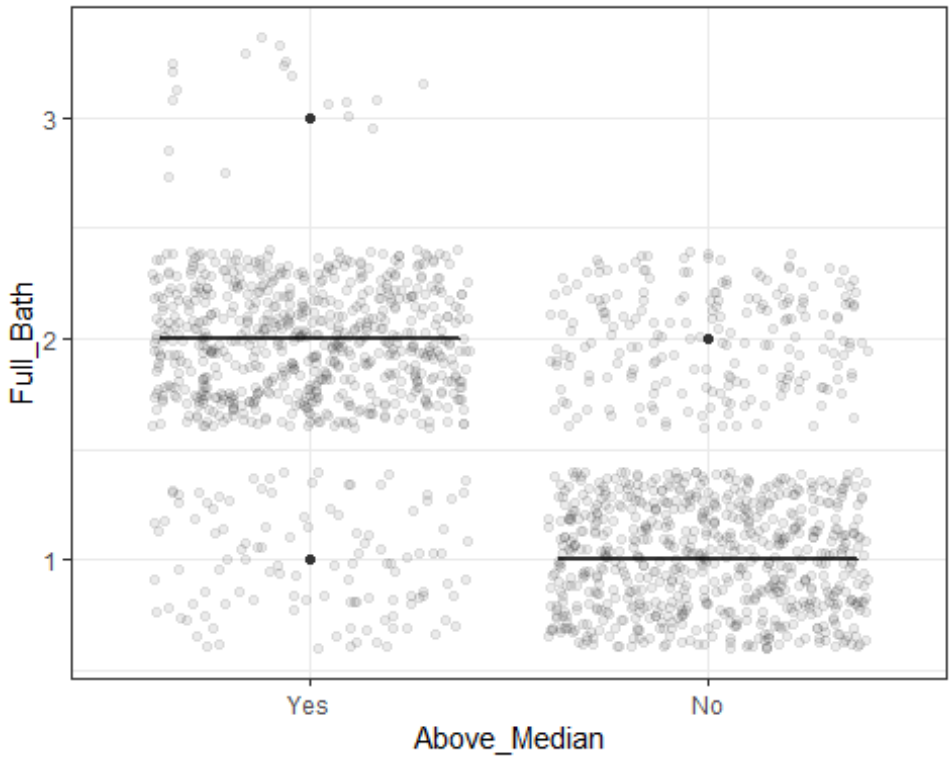

```
ggplot(student, aes(x = Above_Median, y = Gr_Liv_Area)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
```



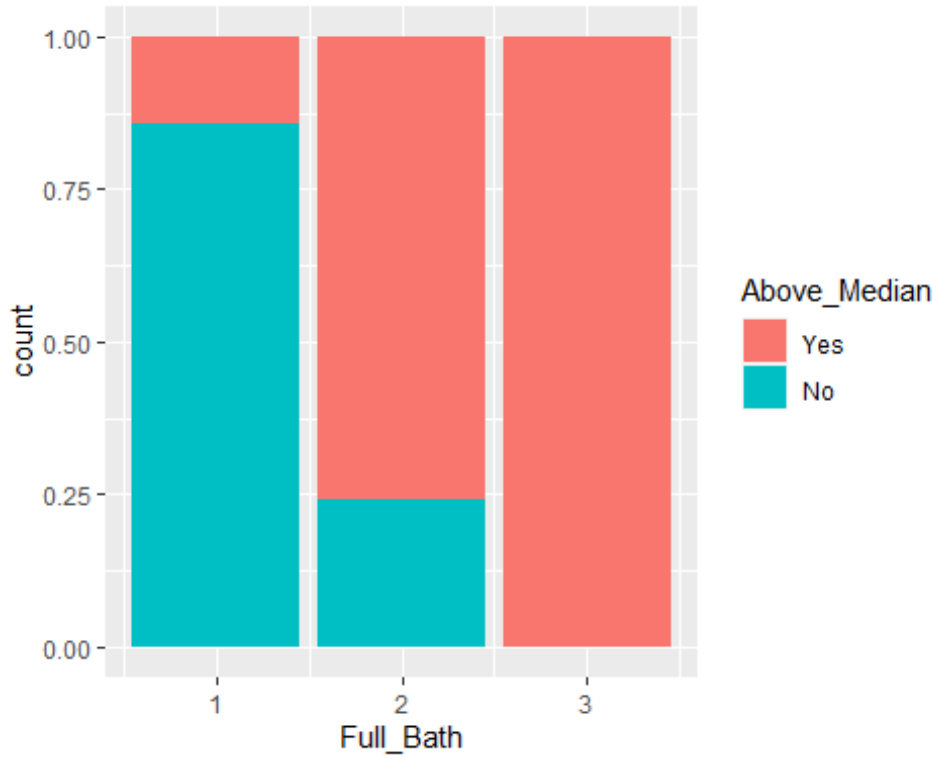

```
ggplot(student, aes(x = Above_Median, y = Year_Built)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
```



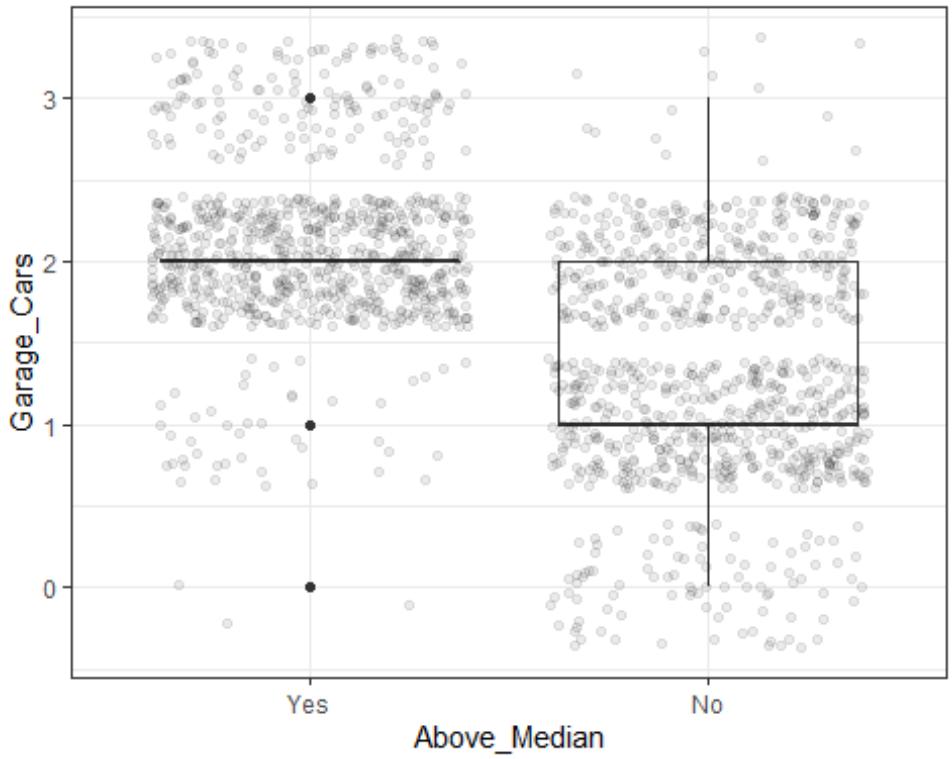
```
ggplot(student, aes(x = Above_Median, y = Full_Bath)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
```



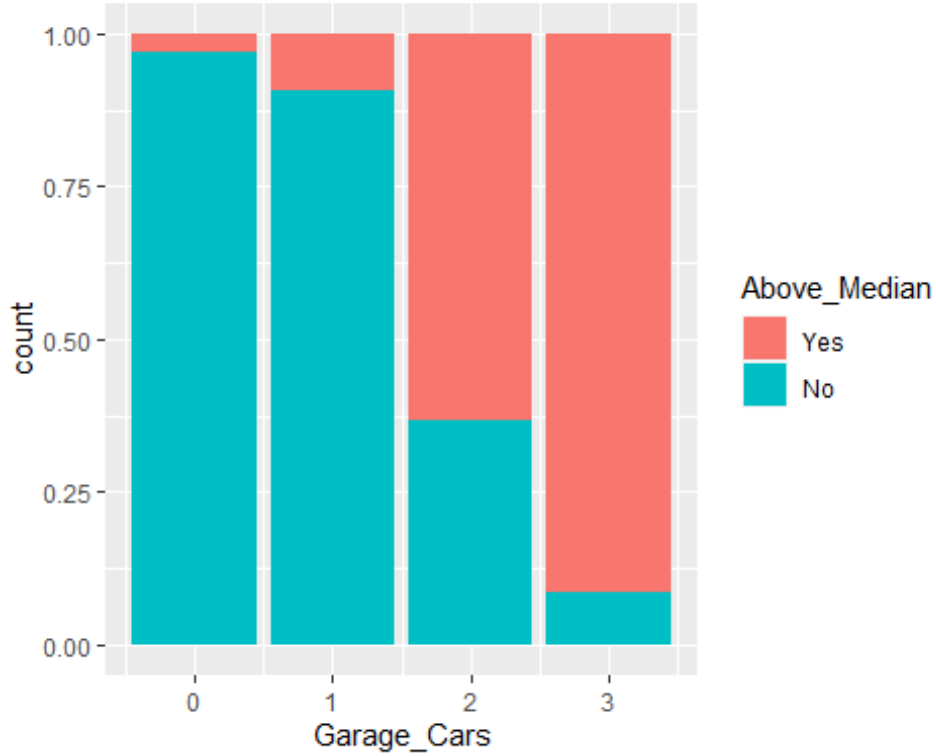

```
ggplot(student, aes(x = Full_Bath, fill = Above_Median)) + geom_bar(position
= "fill")
```




```
t3 = table(student$Above_Median, student$Full_Bath)
prop.table(t3, margin = 2)

##
##                1         2         3
##   Yes 0.1437579 0.7585421 1.0000000
##   No  0.8562421 0.2414579 0.0000000

ggplot(student, aes(x = Above_Median, y = Garage_Cars)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
```




```
ggplot(student, aes(x = Garage_Cars, fill = Above_Median)) +
geom_bar(position = "fill")
```



```
t4 = table(student$Above_Median, student$Garage_Cars)
prop.table(t4, margin = 2)

##
##                0          1          2          3
##   Yes 0.03061224 0.09224319 0.63340336 0.91411043
##   No  0.96938776 0.90775681 0.36659664 0.08588957

p1 = ggplot(student, aes(x = Above_Median, y = Gr_Liv_Area)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p2 = ggplot(student, aes(x = Above_Median, y = Year_Built)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p3 = ggplot(student, aes(x = Above_Median, y = Garage_Cars)) + geom_boxplot()
+ geom_jitter(alpha = 0.08) + theme_bw()
p4 = ggplot(student, aes(x = Above_Median, y = Full_Bath)) + geom_boxplot() +
geom_jitter(alpha = 0.08) + theme_bw()
grid.arrange(p1,p2,p3,p4)
```