#### Linnea Williams - Course Project, Phase 1

ames\_sales <- read\_csv("ames\_student.csv")

## Warning: Missing column names filled in: 'X1' [1]

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## .default = col\_character(),  
## X1 = col\_double(),  
## Lot\_Frontage = col\_double(),  
## Lot\_Area = col\_double(),  
## Year\_Built = col\_double(),  
## Year\_Remod\_Add = col\_double(),  
## Mas\_Vnr\_Area = col\_double(),  
## BsmtFin\_SF\_1 = col\_double(),  
## BsmtFin\_SF\_2 = col\_double(),  
## Bsmt\_Unf\_SF = col\_double(),  
## Total\_Bsmt\_SF = col\_double(),  
## First\_Flr\_SF = col\_double(),  
## Second\_Flr\_SF = col\_double(),  
## Low\_Qual\_Fin\_SF = col\_double(),  
## Gr\_Liv\_Area = col\_double(),  
## Bsmt\_Full\_Bath = col\_double(),  
## Bsmt\_Half\_Bath = col\_double(),  
## Full\_Bath = col\_double(),  
## Half\_Bath = col\_double(),  
## Bedroom\_AbvGr = col\_double(),  
## Kitchen\_AbvGr = col\_double()  
## # ... with 15 more columns  
## )  
## i Use `spec()` for the full column specifications.

# str(ames\_sales)  
# summary(ames\_sales)

ames\_sales = ames\_sales %>%   
 mutate(  
 Year\_Built\_Group = case\_when(  
 Year\_Built <= 1890 ~ "< 1890",  
 Year\_Built > 1890 & Year\_Built <= 1905 ~ "1891-1905",  
 Year\_Built > 1905 & Year\_Built <= 1920 ~ "1906-1920",  
 Year\_Built > 1920 & Year\_Built <= 1935 ~ "1921-1935",  
 Year\_Built > 1935 & Year\_Built <= 1950 ~ "1936-1950",  
 Year\_Built > 1950 & Year\_Built <= 1965 ~ "1951-1965",  
 Year\_Built > 1965 & Year\_Built <= 1980 ~ "1966-1980",  
 Year\_Built > 1980 & Year\_Built <= 1995 ~ "1981-1995",  
 Year\_Built > 1995 & Year\_Built <= 2010 ~ "1996-2010",  
 TRUE ~ "> 2010"  
 )  
 )  
  
ames\_sales = ames\_sales %>%   
 mutate(  
 Year\_RemodAdd\_Group = case\_when(  
 Year\_Remod\_Add <= 1950 ~ "< 1950",  
 Year\_Remod\_Add > 1950 & Year\_Remod\_Add <= 1960 ~ "1951-1960",  
 Year\_Remod\_Add > 1960 & Year\_Remod\_Add <= 1970 ~ "1961-1970",  
 Year\_Remod\_Add > 1970 & Year\_Remod\_Add <= 1980 ~ "1971-1980",  
 Year\_Remod\_Add > 1980 & Year\_Remod\_Add <= 1990 ~ "1981-1990",  
 Year\_Remod\_Add > 1990 & Year\_Remod\_Add <= 2000 ~ "1991-2000",  
 Year\_Remod\_Add > 2000 & Year\_Remod\_Add <= 2010 ~ "2000-2010",  
 TRUE ~ "> 2010"  
 )  
 )  
  
# glimpse(ames\_sales)

ames\_sales <- ames\_sales %>% mutate\_if(is.character, as.factor)  
  
# glimpse(ames\_sales)

# levels(ames\_sales$Lot\_Shape)  
# levels(ames\_sales$Utilities)  
# levels(ames\_sales$Land\_Slope)  
# levels(ames\_sales$Overall\_Qual)  
# levels(ames\_sales$Overall\_Cond)  
# levels(ames\_sales$Exter\_Qual)  
# levels(ames\_sales$Exter\_Cond)  
# levels(ames\_sales$Bsmt\_Qual)  
# levels(ames\_sales$Bsmt\_Cond)  
# levels(ames\_sales$Bsmt\_Exposure)  
# levels(ames\_sales$BsmtFin\_Type\_1)  
# levels(ames\_sales$BsmtFin\_Type\_2)  
# levels(ames\_sales$Heating\_QC)  
# levels(ames\_sales$Electrical)  
# levels(ames\_sales$Kitchen\_Qual)  
# levels(ames\_sales$Functional)  
# levels(ames\_sales$Fireplace\_Qu)  
# levels(ames\_sales$Garage\_Finish)  
# levels(ames\_sales$Garage\_Qual)  
# levels(ames\_sales$Garage\_Cond)  
# levels(ames\_sales$Paved\_Drive)  
# levels(ames\_sales$Pool\_QC)  
# levels(ames\_sales$Fence)  
#levels(ames\_sales$Year\_Built\_Group)  
#levels(ames\_sales$Year\_RemodAdd\_Group)

ames\_sales$Lot\_Shape <-factor(ames\_sales$Lot\_Shape,levels(ames\_sales$Lot\_Shape)  
 [c(3,4,2,1)])  
levels(ames\_sales$Lot\_Shape)

## [1] "Regular" "Slightly\_Irregular" "Moderately\_Irregular"  
## [4] "Irregular"

ames\_sales$Overall\_Qual<-factor(ames\_sales$Overall\_Qual,  
 levels(ames\_sales$Overall\_Qual)  
 [c(8,4,9,6,1,2,3,5,7,10)])  
levels(ames\_sales$Overall\_Qual)

## [1] "Very\_Excellent" "Excellent" "Very\_Good" "Good"   
## [5] "Above\_Average" "Average" "Below\_Average" "Fair"   
## [9] "Poor" "Very\_Poor"

ames\_sales$Overall\_Cond<-factor(ames\_sales$Overall\_Cond,  
 levels(ames\_sales$Overall\_Cond)  
 [c(10,4,8,6,1,2,3,5,7,9)])  
levels(ames\_sales$Overall\_Cond)

## [1] "Excellent" "Very\_Good" "Good" "Above\_Average"  
## [5] "Average" "Below\_Average" "Fair" "Poor"   
## [9] "Very\_Poor"

#Note to self - this data set does not have a "Very Excellent" Overall\_Cond, revisit this on competition data set  
  
ames\_sales$Exter\_Qual<-factor(ames\_sales$Exter\_Qual,  
 levels(ames\_sales$Exter\_Qual)[c(1,3,4,2,5)])  
levels(ames\_sales$Exter\_Qual)

## [1] "Excellent" "Good" "Typical" "Fair"

#Note to self - this data set does not have a "Poor" Exter\_Qual, revisit this on competition data set  
  
ames\_sales$Exter\_Cond <-factor(ames\_sales$Exter\_Cond,  
 levels(ames\_sales$Exter\_Cond)[c(1,3,5,2,4)])  
levels(ames\_sales$Exter\_Cond)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"

ames\_sales$Bsmt\_Qual <-factor(ames\_sales$Bsmt\_Qual,  
 levels(ames\_sales$Bsmt\_Qual)[c(1,3,6,2,5,4)])  
levels(ames\_sales$Bsmt\_Qual)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"   
## [6] "No\_Basement"

ames\_sales$Bsmt\_Cond <-factor(ames\_sales$Bsmt\_Cond,  
 levels(ames\_sales$Bsmt\_Cond)[c(1,3,6,2,5,4)])  
levels(ames\_sales$Bsmt\_Cond)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"   
## [6] "No\_Basement"

ames\_sales$Bsmt\_Exposure <-factor(ames\_sales$Bsmt\_Exposure,  
 levels(ames\_sales$Bsmt\_Exposure)[c(2,1,3,4,5)])  
levels(ames\_sales$Bsmt\_Exposure)

## [1] "Gd" "Av" "Mn" "No" "No\_Basement"

ames\_sales$BsmtFin\_Type\_1 <-factor(ames\_sales$BsmtFin\_Type\_1,  
 levels(ames\_sales$BsmtFin\_Type\_1)  
 [c(3,1,2,6,4,7,5)])  
levels(ames\_sales$BsmtFin\_Type\_1)

## [1] "GLQ" "ALQ" "BLQ" "Rec" "LwQ"   
## [6] "Unf" "No\_Basement"

ames\_sales$BsmtFin\_Type\_2 <-factor(ames\_sales$BsmtFin\_Type\_2,  
 levels(ames\_sales$BsmtFin\_Type\_2)  
 [c(3,1,2,6,4,7,5)])  
levels(ames\_sales$BsmtFin\_Type\_2)

## [1] "GLQ" "ALQ" "BLQ" "Rec" "LwQ"   
## [6] "Unf" "No\_Basement"

ames\_sales$Heating\_QC <-factor(ames\_sales$Heating\_QC,  
 levels(ames\_sales$Heating\_QC)[c(1,3,5,2,4)])  
levels(ames\_sales$Heating\_QC)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"

ames\_sales$Electrical <-factor(ames\_sales$Electrical,  
 levels(ames\_sales$Electrical)[c(4,1,2,3,5)])  
levels(ames\_sales$Electrical)

## [1] "SBrkr" "FuseA" "FuseF" "FuseP" "Unknown"

ames\_sales$Kitchen\_Qual <-factor(ames\_sales$Kitchen\_Qual,  
 levels(ames\_sales$Kitchen\_Qual)[c(1,3,5,2,4)])  
levels(ames\_sales$Kitchen\_Qual)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"

ames\_sales$Functional <-factor(ames\_sales$Functional,  
 levels(ames\_sales$Functional)  
 [c(8,3,4,5,1,2,7,6)])  
levels(ames\_sales$Functional)

## [1] "Typ" "Min1" "Min2" "Mod" "Maj1" "Maj2" "Sev" "Sal"

ames\_sales$Fireplace\_Qu <-factor(ames\_sales$Fireplace\_Qu,  
 levels(ames\_sales$Fireplace\_Qu)[c(1,3,6,2,5,4)])  
levels(ames\_sales$Fireplace\_Qu)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor"   
## [6] "No\_Fireplace"

ames\_sales$Garage\_Finish <-factor(ames\_sales$Garage\_Finish,  
 levels(ames\_sales$Garage\_Finish)  
 [c(1,3,4,2)])  
levels(ames\_sales$Garage\_Finish)

## [1] "Fin" "RFn" "Unf" "No\_Garage"

ames\_sales$Garage\_Qual <-factor(ames\_sales$Garage\_Qual,  
 levels(ames\_sales$Garage\_Qual)[c(1,3,6,2,5,4)])  
levels(ames\_sales$Garage\_Qual)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor" "No\_Garage"

ames\_sales$Garage\_Cond <-factor(ames\_sales$Garage\_Cond,  
 levels(ames\_sales$Garage\_Cond)[c(1,3,6,2,5,4)])  
levels(ames\_sales$Garage\_Cond)

## [1] "Excellent" "Good" "Typical" "Fair" "Poor" "No\_Garage"

ames\_sales$Paved\_Drive <-factor(ames\_sales$Paved\_Drive,  
 levels(ames\_sales$Paved\_Drive)[c(3,2,1)])  
levels(ames\_sales$Paved\_Drive)

## [1] "Paved" "Partial\_Pavement" "Dirt\_Gravel"

ames\_sales$Pool\_QC <-factor(ames\_sales$Pool\_QC,  
 levels(ames\_sales$Pool\_QC)[c(1,3,5,2,4)])  
levels(ames\_sales$Pool\_QC)

## [1] "Excellent" "Good" "Typical" "Fair" "No\_Pool"

ames\_sales$Fence <-factor(ames\_sales$Fence,levels(ames\_sales$Fence)  
 [c(1,3,2,4,5)])  
levels(ames\_sales$Fence)

## [1] "Good\_Privacy" "Minimum\_Privacy" "Good\_Wood"   
## [4] "Minimum\_Wood\_Wire" "No\_Fence"

levels(ames\_sales$Neighborhood)

## [1] "Bloomington\_Heights"   
## [2] "Blueste"   
## [3] "Briardale"   
## [4] "Brookside"   
## [5] "Clear\_Creek"   
## [6] "College\_Creek"   
## [7] "Crawford"   
## [8] "Edwards"   
## [9] "Gilbert"   
## [10] "Green\_Hills"   
## [11] "Greens"   
## [12] "Iowa\_DOT\_and\_Rail\_Road"   
## [13] "Landmark"   
## [14] "Meadow\_Village"   
## [15] "Mitchell"   
## [16] "North\_Ames"   
## [17] "Northpark\_Villa"   
## [18] "Northridge"   
## [19] "Northridge\_Heights"   
## [20] "Northwest\_Ames"   
## [21] "Old\_Town"   
## [22] "Sawyer"   
## [23] "Sawyer\_West"   
## [24] "Somerset"   
## [25] "South\_and\_West\_of\_Iowa\_State\_University"  
## [26] "Stone\_Brook"   
## [27] "Timberland"   
## [28] "Veenker"

# Had a duplicate Overall\_Qual code that wrote over the first code, now to fix again  
  
# ames\_sales$Overall\_Qual<-factor(ames\_sales$Overall\_Qual,  
# levels(ames\_sales$Overall\_Qual)  
# [c(5,6,7,2,8,4,9,1,3,10)])  
levels(ames\_sales$Overall\_Qual)

## [1] "Very\_Excellent" "Excellent" "Very\_Good" "Good"   
## [5] "Above\_Average" "Average" "Below\_Average" "Fair"   
## [9] "Poor" "Very\_Poor"

ames\_sales <- ames\_sales %>%   
 mutate(Bsmt\_Full\_Bath = as\_factor(Bsmt\_Full\_Bath)) %>%   
 mutate(Bsmt\_Half\_Bath = as\_factor(Bsmt\_Half\_Bath)) %>%   
 mutate(Full\_Bath = as\_factor(Full\_Bath)) %>%   
 mutate(Half\_Bath = as\_factor(Half\_Bath)) %>%   
 mutate(Bedroom\_AbvGr = as\_factor(Bedroom\_AbvGr)) %>%   
 mutate(Kitchen\_AbvGr = as\_factor(Kitchen\_AbvGr)) %>%   
 mutate(TotRms\_AbvGrd = as\_factor(TotRms\_AbvGrd)) %>%   
 mutate(Fireplaces = as\_factor(Fireplaces)) %>%   
 mutate(Garage\_Cars = as\_factor(Garage\_Cars))%>%  
   
 mutate(Mo\_Sold = as\_factor(Mo\_Sold)) %>%   
 mutate(Year\_Sold = as\_factor(Year\_Sold))   
  
#glimpse(ames\_sales)

ames\_sales = ames\_sales %>%   
 mutate(Land\_Slope = fct\_recode(Land\_Slope, "Gentle" = "Gtl",   
 "Moderate" = "Mod", "Severe" = "Sev" )) %>%  
 mutate(Bsmt\_Exposure = fct\_recode(Bsmt\_Exposure, "Good" = "Gd",   
 "Average" = "Av", "Minimum" = "Mn",   
 "No Exposure" = "No",   
 "No Basement" = "NA" )) %>%  
 mutate(BsmtFin\_Type\_1 = fct\_recode(BsmtFin\_Type\_1, "Good Living" = "GLQ",   
 "Avg Living" = "ALQ",   
 "Below Avg Living" ="BLQ",   
 "Avg Rec" = "Rec",   
 "Low Quality" = "LwQ",  
 "Unfinished" = "Unf", "No Basement" = "NA"))%>%  
 mutate(BsmtFin\_Type\_2 = fct\_recode(BsmtFin\_Type\_2, "Good Living" = "GLQ",   
 "Avg Living" = "ALQ",   
 "Below Avg Living" ="BLQ",   
 "Avg Rec" = "Rec",   
 "Low Quality" = "LwQ",  
 "Unfinished" = "Unf", "No Basement" = "NA"))%>%  
 mutate(Central\_Air = fct\_recode(Central\_Air, "Yes" = "Y",   
 "No" = "N")) %>%  
 mutate(Garage\_Finish = fct\_recode(Garage\_Finish, "Finished" = "Fin",   
 "Rough Finished" = "RFn",   
 "Unfinished" = "Unf", "No Garge" = "NA"))

## Warning: Unknown levels in `f`: NA  
  
## Warning: Unknown levels in `f`: NA  
  
## Warning: Unknown levels in `f`: NA  
  
## Warning: Unknown levels in `f`: NA

ames\_sales = ames\_sales %>%   
 mutate(Neighborhood = fct\_recode(Neighborhood, "S and W of ISU" =   
 "South\_and\_West\_of\_Iowa\_State\_University")  
 )  
  
  
levels(ames\_sales$Garage\_Finish)

## [1] "Finished" "Rough Finished" "Unfinished" "No\_Garage"

#glimpse(ames\_sales)

skim(ames\_sales)

Data summary

|  |  |
| --- | --- |
| Name | ames\_sales |
| Number of rows | 2053 |
| Number of columns | 84 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 60 |
| numeric | 24 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| MS\_SubClass | 0 | 1 | FALSE | 16 | One: 772, Two: 383, One: 204, One: 129 |
| MS\_Zoning | 0 | 1 | FALSE | 7 | Res: 1600, Res: 326, Flo: 87, Res: 20 |
| Street | 0 | 1 | FALSE | 2 | Pav: 2046, Grv: 7 |
| Alley | 0 | 1 | FALSE | 3 | No\_: 1914, Gra: 94, Pav: 45 |
| Lot\_Shape | 0 | 1 | FALSE | 4 | Reg: 1275, Sli: 714, Mod: 53, Irr: 11 |
| Land\_Contour | 0 | 1 | FALSE | 4 | Lvl: 1833, HLS: 94, Bnk: 81, Low: 45 |
| Utilities | 0 | 1 | FALSE | 2 | All: 2052, NoS: 1 |
| Lot\_Config | 0 | 1 | FALSE | 5 | Ins: 1495, Cor: 359, Cul: 135, FR2: 56 |
| Land\_Slope | 0 | 1 | FALSE | 3 | Gen: 1951, Mod: 89, Sev: 13 |
| Neighborhood | 0 | 1 | FALSE | 28 | Nor: 327, Col: 183, Old: 181, Edw: 129 |
| Condition\_1 | 0 | 1 | FALSE | 9 | Nor: 1771, Fee: 113, Art: 67, RRA: 35 |
| Condition\_2 | 0 | 1 | FALSE | 8 | Nor: 2027, Fee: 12, Art: 4, Pos: 4 |
| Bldg\_Type | 0 | 1 | FALSE | 5 | One: 1706, Twn: 157, Dup: 76, Twn: 67 |
| House\_Style | 0 | 1 | FALSE | 8 | One: 1052, Two: 590, One: 225, SLv: 90 |
| Overall\_Qual | 0 | 1 | FALSE | 10 | Ave: 587, Abo: 518, Goo: 411, Ver: 237 |
| Overall\_Cond | 0 | 1 | FALSE | 9 | Ave: 1143, Abo: 376, Goo: 286, Ver: 98 |
| Roof\_Style | 0 | 1 | FALSE | 6 | Gab: 1607, Hip: 404, Fla: 14, Gam: 14 |
| Roof\_Matl | 0 | 1 | FALSE | 6 | Com: 2023, Tar: 17, WdS: 8, WdS: 3 |
| Exterior\_1st | 0 | 1 | FALSE | 16 | Vin: 705, Met: 319, Wd : 313, HdB: 303 |
| Exterior\_2nd | 0 | 1 | FALSE | 17 | Vin: 699, Met: 317, Wd : 302, HdB: 277 |
| Mas\_Vnr\_Type | 0 | 1 | FALSE | 5 | Non: 1231, Brk: 638, Sto: 166, Brk: 17 |
| Exter\_Qual | 0 | 1 | FALSE | 4 | Typ: 1272, Goo: 682, Exc: 78, Fai: 21 |
| Exter\_Cond | 0 | 1 | FALSE | 5 | Typ: 1787, Goo: 213, Fai: 43, Exc: 9 |
| Foundation | 0 | 1 | FALSE | 6 | PCo: 911, CBl: 880, Brk: 216, Sla: 36 |
| Bsmt\_Qual | 0 | 1 | FALSE | 6 | Typ: 911, Goo: 849, Exc: 178, Fai: 57 |
| Bsmt\_Cond | 0 | 1 | FALSE | 6 | Typ: 1833, Goo: 80, Fai: 76, No\_: 57 |
| Bsmt\_Exposure | 0 | 1 | FALSE | 5 | No : 1331, Ave: 284, Goo: 199, Min: 179 |
| BsmtFin\_Type\_1 | 0 | 1 | FALSE | 7 | Unf: 602, Goo: 578, Avg: 298, Avg: 216 |
| BsmtFin\_Type\_2 | 0 | 1 | FALSE | 7 | Unf: 1740, Avg: 79, Low: 64, No\_: 58 |
| Heating | 0 | 1 | FALSE | 6 | Gas: 2019, Gas: 21, Gra: 6, Wal: 5 |
| Heating\_QC | 0 | 1 | FALSE | 5 | Exc: 1040, Typ: 618, Goo: 333, Fai: 61 |
| Central\_Air | 0 | 1 | FALSE | 2 | Yes: 1916, No: 137 |
| Electrical | 0 | 1 | FALSE | 5 | SBr: 1887, Fus: 126, Fus: 33, Fus: 6 |
| Bsmt\_Full\_Bath | 0 | 1 | FALSE | 4 | 0: 1201, 1: 823, 2: 27, 3: 2 |
| Bsmt\_Half\_Bath | 0 | 1 | FALSE | 3 | 0: 1936, 1: 115, 2: 2 |
| Full\_Bath | 0 | 1 | FALSE | 5 | 2: 1080, 1: 920, 3: 41, 0: 10 |
| Half\_Bath | 0 | 1 | FALSE | 3 | 0: 1300, 1: 736, 2: 17 |
| Bedroom\_AbvGr | 0 | 1 | FALSE | 7 | 3: 1105, 2: 527, 4: 297, 1: 73 |
| Kitchen\_AbvGr | 0 | 1 | FALSE | 3 | 1: 1959, 2: 92, 3: 2 |
| Kitchen\_Qual | 0 | 1 | FALSE | 5 | Typ: 1070, Goo: 790, Exc: 142, Fai: 50 |
| TotRms\_AbvGrd | 0 | 1 | FALSE | 11 | 6: 589, 7: 468, 5: 416, 8: 245 |
| Functional | 0 | 1 | FALSE | 8 | Typ: 1896, Min: 54, Min: 51, Mod: 27 |
| Fireplaces | 0 | 1 | FALSE | 5 | 0: 993, 1: 891, 2: 161, 3: 7 |
| Fireplace\_Qu | 0 | 1 | FALSE | 6 | No\_: 993, Goo: 538, Typ: 409, Fai: 56 |
| Garage\_Type | 0 | 1 | FALSE | 7 | Att: 1204, Det: 549, Bui: 127, No\_: 108 |
| Garage\_Finish | 0 | 1 | FALSE | 4 | Unf: 872, Rou: 563, Fin: 509, No\_: 109 |
| Garage\_Cars | 0 | 1 | FALSE | 6 | 2: 1131, 1: 539, 3: 261, 0: 108 |
| Garage\_Qual | 0 | 1 | FALSE | 6 | Typ: 1839, No\_: 109, Fai: 85, Goo: 16 |
| Garage\_Cond | 0 | 1 | FALSE | 6 | Typ: 1872, No\_: 109, Fai: 53, Goo: 10 |
| Paved\_Drive | 0 | 1 | FALSE | 3 | Pav: 1848, Dir: 163, Par: 42 |
| Pool\_QC | 0 | 1 | FALSE | 5 | No\_: 2047, Exc: 2, Typ: 2, Goo: 1 |
| Fence | 0 | 1 | FALSE | 5 | No\_: 1661, Min: 225, Goo: 81, Goo: 77 |
| Misc\_Feature | 0 | 1 | FALSE | 5 | Non: 1978, She: 66, Gar: 5, Oth: 3 |
| Mo\_Sold | 0 | 1 | FALSE | 12 | 6: 352, 7: 320, 5: 275, 4: 187 |
| Year\_Sold | 0 | 1 | FALSE | 5 | 200: 499, 200: 456, 200: 445, 200: 442 |
| Sale\_Type | 0 | 1 | FALSE | 10 | WD: 1789, New: 163, COD: 54, Con: 16 |
| Sale\_Condition | 0 | 1 | FALSE | 6 | Nor: 1712, Par: 169, Abn: 121, Fam: 30 |
| Above\_Median | 0 | 1 | FALSE | 2 | Yes: 1043, No: 1010 |
| Year\_Built\_Group | 0 | 1 | FALSE | 9 | 199: 655, 196: 381, 195: 369, 198: 167 |
| Year\_RemodAdd\_Group | 0 | 1 | FALSE | 7 | 200: 680, 199: 399, < 1: 256, 197: 225 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| X1 | 0 | 1 | 1027.00 | 592.79 | 1.00 | 514.00 | 1027.00 | 1540.00 | 2053.00 | ▇▇▇▇▇ |
| Lot\_Frontage | 0 | 1 | 57.38 | 33.20 | 0.00 | 43.00 | 62.00 | 78.00 | 313.00 | ▇▇▁▁▁ |
| Lot\_Area | 0 | 1 | 10258.40 | 8427.38 | 1300.00 | 7500.00 | 9548.00 | 11600.00 | 215245.00 | ▇▁▁▁▁ |
| Year\_Built | 0 | 1 | 1970.64 | 30.40 | 1875.00 | 1953.00 | 1972.00 | 2000.00 | 2010.00 | ▁▂▃▆▇ |
| Year\_Remod\_Add | 0 | 1 | 1984.08 | 20.96 | 1950.00 | 1965.00 | 1993.00 | 2004.00 | 2010.00 | ▅▂▂▃▇ |
| Mas\_Vnr\_Area | 0 | 1 | 103.75 | 183.59 | 0.00 | 0.00 | 0.00 | 164.00 | 1600.00 | ▇▁▁▁▁ |
| BsmtFin\_SF\_1 | 0 | 1 | 4.21 | 2.24 | 1.00 | 3.00 | 3.00 | 7.00 | 7.00 | ▅▆▁▁▇ |
| BsmtFin\_SF\_2 | 0 | 1 | 52.57 | 175.99 | 0.00 | 0.00 | 0.00 | 0.00 | 1526.00 | ▇▁▁▁▁ |
| Bsmt\_Unf\_SF | 0 | 1 | 561.19 | 441.72 | 0.00 | 226.00 | 460.00 | 801.00 | 2336.00 | ▇▅▂▁▁ |
| Total\_Bsmt\_SF | 0 | 1 | 1054.57 | 435.33 | 0.00 | 793.00 | 988.00 | 1304.00 | 5095.00 | ▇▇▁▁▁ |
| First\_Flr\_SF | 0 | 1 | 1167.52 | 391.79 | 432.00 | 882.00 | 1088.00 | 1402.00 | 5095.00 | ▇▃▁▁▁ |
| Second\_Flr\_SF | 0 | 1 | 326.07 | 422.44 | 0.00 | 0.00 | 0.00 | 701.00 | 1862.00 | ▇▂▂▁▁ |
| Low\_Qual\_Fin\_SF | 0 | 1 | 4.97 | 49.09 | 0.00 | 0.00 | 0.00 | 0.00 | 1064.00 | ▇▁▁▁▁ |
| Gr\_Liv\_Area | 0 | 1 | 1498.56 | 487.84 | 480.00 | 1137.00 | 1447.00 | 1737.00 | 5095.00 | ▇▇▁▁▁ |
| Garage\_Area | 0 | 1 | 471.96 | 213.43 | 0.00 | 320.00 | 478.00 | 576.00 | 1488.00 | ▃▇▂▁▁ |
| Wood\_Deck\_SF | 0 | 1 | 93.52 | 127.71 | 0.00 | 0.00 | 0.00 | 168.00 | 1424.00 | ▇▁▁▁▁ |
| Open\_Porch\_SF | 0 | 1 | 48.17 | 69.51 | 0.00 | 0.00 | 27.00 | 72.00 | 742.00 | ▇▁▁▁▁ |
| Enclosed\_Porch | 0 | 1 | 23.02 | 60.59 | 0.00 | 0.00 | 0.00 | 0.00 | 584.00 | ▇▁▁▁▁ |
| Three\_season\_porch | 0 | 1 | 2.80 | 25.65 | 0.00 | 0.00 | 0.00 | 0.00 | 407.00 | ▇▁▁▁▁ |
| Screen\_Porch | 0 | 1 | 16.68 | 57.94 | 0.00 | 0.00 | 0.00 | 0.00 | 576.00 | ▇▁▁▁▁ |
| Pool\_Area | 0 | 1 | 1.34 | 27.74 | 0.00 | 0.00 | 0.00 | 0.00 | 800.00 | ▇▁▁▁▁ |
| Misc\_Val | 0 | 1 | 60.12 | 662.76 | 0.00 | 0.00 | 0.00 | 0.00 | 17000.00 | ▇▁▁▁▁ |
| Longitude | 0 | 1 | -93.64 | 0.03 | -93.69 | -93.66 | -93.64 | -93.62 | -93.58 | ▅▅▇▇▁ |
| Latitude | 0 | 1 | 42.03 | 0.02 | 41.99 | 42.02 | 42.03 | 42.05 | 42.06 | ▂▂▇▇▇ |

# ames\_sales %>% count(Bsmt\_Full\_Bath)  
# ames\_sales %>% count(Bsmt\_Half\_Bath)  
# ames\_sales %>% count(Full\_Bath)  
# ames\_sales %>% count(Half\_Bath)  
# ames\_sales %>% count(Bedroom\_AbvGr)  
# ames\_sales %>% count(Kitchen\_AbvGr)  
# ames\_sales %>% count(TotRms\_AbvGrd)  
# ames\_sales %>% count(Fireplaces)  
# ames\_sales %>% count(Garage\_Cars)  
#   
# I don't think these would be good predictors just based on the majority of observations equaling 0  
  
# ames\_sales %>% count(Misc\_Feature)  
# ames\_sales %>% count(Pool\_QC)  
# ames\_sales %>% count(Fence)  
# ames\_sales %>% count(Pool\_Area)  
#   
# My initial thought is that these two columns serve the same purpose as Neighborhood  
  
# ames\_sales %>% count(Latitude)  
# ames\_sales %>% count(Longitude)  
#   
# Maybe these should be changed to Yes/No factors instead  
  
# ames\_sales %>% count(Mas\_Vnr\_Area)  
# ames\_sales %>% count(BsmtFin\_SF\_2)  
# ames\_sales %>% count(Second\_Flr\_SF)  
# ames\_sales %>% count(Low\_Qual\_Fin\_SF)  
# ames\_sales %>% count(Open\_Porch\_SF)  
# ames\_sales %>% count(Enclosed\_Porch)  
# ames\_sales %>% count(Three\_season\_porch)  
# ames\_sales %>% count(Screen\_Porch)  
  
  
# Electrical had a an "Unknown" value when leveling - to remove  
  
# ames\_sales %>% count(Electrical)  
#   
# ames\_sales %>% count(Neighborhood, sort = TRUE)  
# ames\_sales %>% count(Neighborhood, Above\_Median, sort = TRUE)  
#   
# ames\_salesClean %>% count(Year\_RemodAdd\_Group, Above\_Median, sort = TRUE)

colSums(ames\_sales != 0)

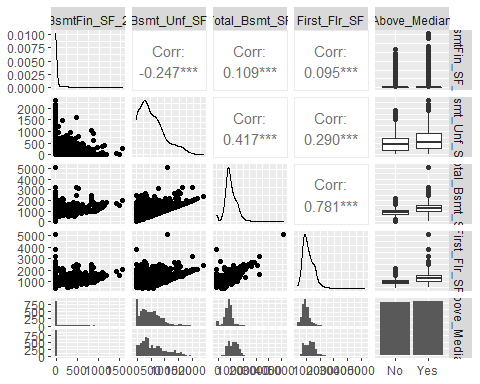
## X1 MS\_SubClass MS\_Zoning Lot\_Frontage   
## 2053 2053 2053 1704   
## Lot\_Area Street Alley Lot\_Shape   
## 2053 2053 2053 2053   
## Land\_Contour Utilities Lot\_Config Land\_Slope   
## 2053 2053 2053 2053   
## Neighborhood Condition\_1 Condition\_2 Bldg\_Type   
## 2053 2053 2053 2053   
## House\_Style Overall\_Qual Overall\_Cond Year\_Built   
## 2053 2053 2053 2053   
## Year\_Remod\_Add Roof\_Style Roof\_Matl Exterior\_1st   
## 2053 2053 2053 2053   
## Exterior\_2nd Mas\_Vnr\_Type Mas\_Vnr\_Area Exter\_Qual   
## 2053 2053 826 2053   
## Exter\_Cond Foundation Bsmt\_Qual Bsmt\_Cond   
## 2053 2053 2053 2053   
## Bsmt\_Exposure BsmtFin\_Type\_1 BsmtFin\_SF\_1 BsmtFin\_Type\_2   
## 2053 2053 2053 2053   
## BsmtFin\_SF\_2 Bsmt\_Unf\_SF Total\_Bsmt\_SF Heating   
## 255 1879 1996 2053   
## Heating\_QC Central\_Air Electrical First\_Flr\_SF   
## 2053 2053 2053 2053   
## Second\_Flr\_SF Low\_Qual\_Fin\_SF Gr\_Liv\_Area Bsmt\_Full\_Bath   
## 861 29 2053 852   
## Bsmt\_Half\_Bath Full\_Bath Half\_Bath Bedroom\_AbvGr   
## 117 2043 753 2046   
## Kitchen\_AbvGr Kitchen\_Qual TotRms\_AbvGrd Functional   
## 2053 2053 2053 2053   
## Fireplaces Fireplace\_Qu Garage\_Type Garage\_Finish   
## 1060 2053 2053 2053   
## Garage\_Cars Garage\_Area Garage\_Qual Garage\_Cond   
## 1945 1945 2053 2053   
## Paved\_Drive Wood\_Deck\_SF Open\_Porch\_SF Enclosed\_Porch   
## 2053 969 1146 337   
## Three\_season\_porch Screen\_Porch Pool\_Area Pool\_QC   
## 28 185 6 2053   
## Fence Misc\_Feature Misc\_Val Mo\_Sold   
## 2053 2053 74 2053   
## Year\_Sold Sale\_Type Sale\_Condition Longitude   
## 2053 2053 2053 2053   
## Latitude Above\_Median Year\_Built\_Group Year\_RemodAdd\_Group   
## 2053 2053 2053 2053

ames\_salesClean = ames\_sales[!grepl("Unknown",ames\_sales$Electrical),]  
  
ames\_salesClean %>% count(Electrical)

## # A tibble: 4 x 2  
## Electrical n  
## \* <fct> <int>  
## 1 SBrkr 1887  
## 2 FuseA 126  
## 3 FuseF 33  
## 4 FuseP 6

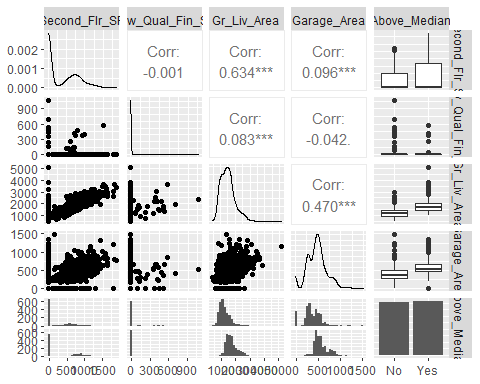
# ggpairs(ames\_salesClean, columns = c(4:5, 27, 35, 82))  
ggpairs(ames\_salesClean, columns = c(37:39, 44, 82))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggpairs(ames\_salesClean, columns = c(45:47, 62, 82))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

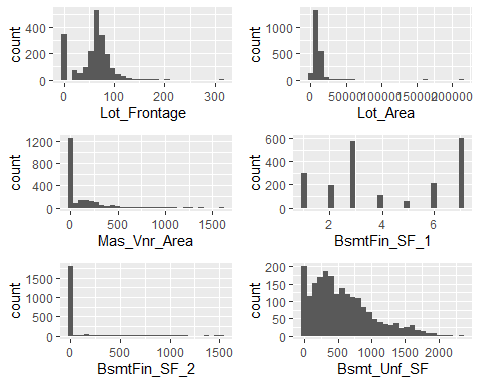


# ggpairs(ames\_salesClean, columns = c(66:69, 82))  
# ggpairs(ames\_salesClean, columns = c(70:71, 75, 80:82))  
  
# ggpairs(ames\_salesClean, columns = c(48:52,82))  
# ggpairs(ames\_salesClean, columns = c(53,55,57,61,82))

Decent correlation (0.781) between First\_Flr\_SF and Total\_Bsmt\_SF. Possible competing predictor variables. If both of these prove to be strong predictors of Above\_Median.

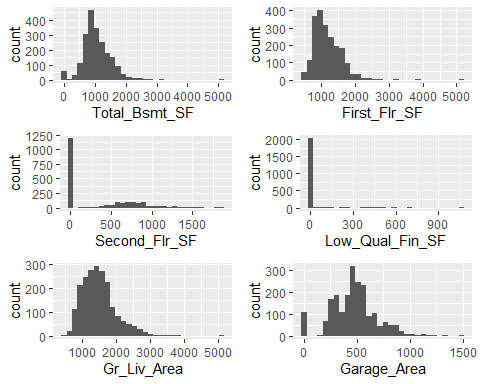
h1 = ggplot(ames\_salesClean, aes(x=Lot\_Frontage))+  
 geom\_histogram()  
h2 = ggplot(ames\_salesClean, aes(x=Lot\_Area))+  
 geom\_histogram()  
h3 = ggplot(ames\_salesClean, aes(x=Mas\_Vnr\_Area))+  
 geom\_histogram()  
h4 = ggplot(ames\_salesClean, aes(x=BsmtFin\_SF\_1))+  
 geom\_histogram()  
h5 = ggplot(ames\_salesClean, aes(x=BsmtFin\_SF\_2))+  
 geom\_histogram()  
h6 = ggplot(ames\_salesClean, aes(x=Bsmt\_Unf\_SF))+  
 geom\_histogram()  
h7 = ggplot(ames\_salesClean, aes(x=Total\_Bsmt\_SF))+  
 geom\_histogram()  
h8 = ggplot(ames\_salesClean, aes(x=First\_Flr\_SF))+  
 geom\_histogram()  
h9 = ggplot(ames\_salesClean, aes(x=Second\_Flr\_SF))+  
 geom\_histogram()  
h10 = ggplot(ames\_salesClean, aes(x=Low\_Qual\_Fin\_SF))+  
 geom\_histogram()  
h11 = ggplot(ames\_salesClean, aes(x=Gr\_Liv\_Area))+  
 geom\_histogram()  
h12 = ggplot(ames\_salesClean, aes(x=Garage\_Area))+  
 geom\_histogram()  
h13 = ggplot(ames\_salesClean, aes(x=Wood\_Deck\_SF))+  
 geom\_histogram()  
h14 = ggplot(ames\_salesClean, aes(x=Open\_Porch\_SF))+  
 geom\_histogram()  
h15 = ggplot(ames\_salesClean, aes(x=Enclosed\_Porch))+  
 geom\_histogram()  
h16 = ggplot(ames\_salesClean, aes(x=Three\_season\_porch))+  
 geom\_histogram()  
h17 = ggplot(ames\_salesClean, aes(x=Screen\_Porch))+  
 geom\_histogram()  
h18 = ggplot(ames\_salesClean, aes(x=Pool\_Area))+  
 geom\_histogram()  
h19 = ggplot(ames\_salesClean, aes(x=Misc\_Val))+  
 geom\_histogram()  
h20 = ggplot(ames\_salesClean, aes(x=Longitude))+  
 geom\_histogram()  
h21 = ggplot(ames\_salesClean, aes(x=Latitude))+  
 geom\_histogram()  
  
grid.arrange(h1,h2,h3,h4,h5,h6,ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



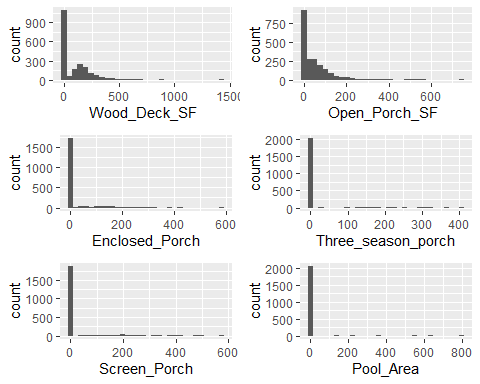
grid.arrange(h7,h8,h9,h10,h11,h12,ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



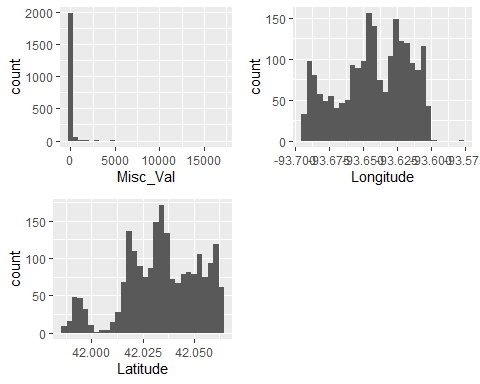
grid.arrange(h13,h14,h15,h16,h17,h18,ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



grid.arrange(h19,h20,h21, ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

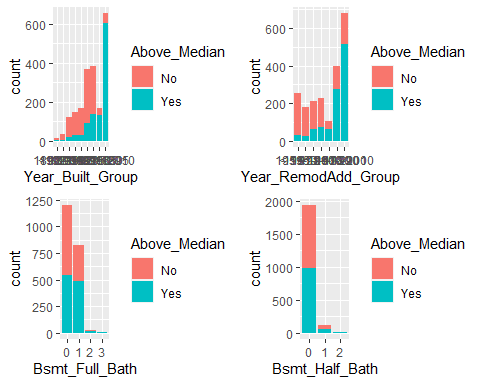


#Many of the Continuous variables have large amount of "o" observations.

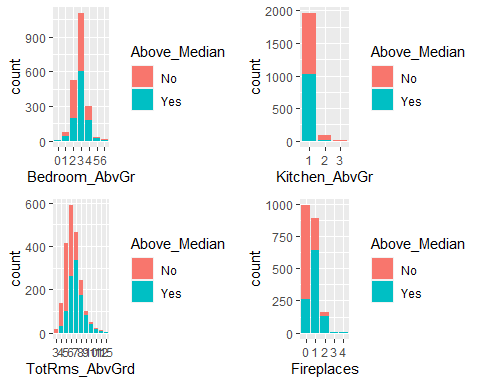
Many of the continuous variables have most of their observations listed as 0. Eliminating these variables as possible predictors for Above\_Median (would recommend removing these variables entirely from the data set).

Will look into Total\_Bsmt\_SF, First\_FLr\_SF,Gr\_Liv\_Area, Garage\_Area, Lot Frontage, Bsmt\_Fin\_SF\_1, BSmt\_Unf\_SF further as predictors of Above\_Median as Box Plots.

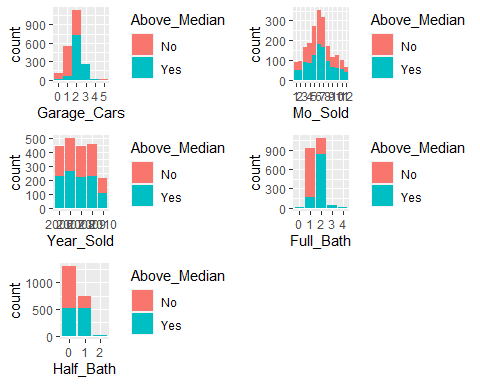
pd1 = ggplot(ames\_salesClean, aes(x=Year\_Built\_Group, fill = Above\_Median)) +   
 geom\_bar()  
pd2 = ggplot(ames\_salesClean, aes(x=Year\_RemodAdd\_Group, fill = Above\_Median)) +   
 geom\_bar()  
pd3 = ggplot(ames\_salesClean, aes(x=Bsmt\_Full\_Bath, fill = Above\_Median)) +   
 geom\_bar()  
pd4 = ggplot(ames\_salesClean, aes(x=Bsmt\_Half\_Bath, fill = Above\_Median)) +   
 geom\_bar()  
pd5 = ggplot(ames\_salesClean, aes(x=Bedroom\_AbvGr, fill = Above\_Median)) +   
 geom\_bar()  
pd6 = ggplot(ames\_salesClean, aes(x=Kitchen\_AbvGr, fill = Above\_Median)) +   
 geom\_bar()  
pd7 = ggplot(ames\_salesClean, aes(x=TotRms\_AbvGrd, fill = Above\_Median)) +   
 geom\_bar()  
pd8 = ggplot(ames\_salesClean, aes(x=Fireplaces, fill = Above\_Median)) +  
 geom\_bar()  
pd9 = ggplot(ames\_salesClean, aes(x=Garage\_Cars, fill = Above\_Median)) +   
 geom\_bar()  
pd10 = ggplot(ames\_salesClean, aes(x=Mo\_Sold, fill = Above\_Median)) +   
 geom\_bar()  
pd11 = ggplot(ames\_salesClean, aes(x=Year\_Sold, fill = Above\_Median)) +   
 geom\_bar()  
pd12 = ggplot(ames\_salesClean, aes(x=Full\_Bath, fill = Above\_Median)) +   
 geom\_bar()  
pd13 = ggplot(ames\_salesClean, aes(x=Half\_Bath, fill = Above\_Median)) +   
 geom\_bar()  
  
  
grid.arrange(pd1,pd2,pd3,pd4,ncol=2)



grid.arrange(pd5,pd6,pd7,pd8,ncol=2)



grid.arrange(pd9,pd10,pd11,pd12,pd13,ncol=2)



Will further explore Bedroom\_AbvGr, TotalRms\_AbvGrd, Fireplaces, Garage\_Cars, & Full\_Bath a little further as predictors.

All other variables in this group appear to have equally distributed observations of Yes and No across the different levels.

pn24 = ggplot(ames\_salesClean, aes(x=MS\_SubClass, fill = Above\_Median))+  
 geom\_bar()+  
 theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))  
pn25 = ggplot(ames\_salesClean, aes(x=MS\_Zoning, fill = Above\_Median)) +  
 geom\_bar()  
pn26 = ggplot(ames\_salesClean, aes(x=Street, fill = Above\_Median)) +  
 geom\_bar()  
pn27 = ggplot(ames\_salesClean, aes(x=Alley, fill = Above\_Median)) +  
 geom\_bar()  
pn28 = ggplot(ames\_salesClean, aes(x=Land\_Contour, fill = Above\_Median)) +  
 geom\_bar()  
pn29 = ggplot(ames\_salesClean, aes(x=Lot\_Config, fill = Above\_Median)) +  
 geom\_bar()  
pn30 = ggplot(ames\_salesClean, aes(x=Neighborhood,fill = Above\_Median)) +  
 geom\_bar()  
pn31 = ggplot(ames\_salesClean, aes(x=Condition\_1, fill = Above\_Median)) +  
 geom\_bar()  
pn32 = ggplot(ames\_salesClean, aes(x=Condition\_2, fill = Above\_Median)) +  
 geom\_bar()  
pn33 = ggplot(ames\_salesClean, aes(x=Bldg\_Type, fill = Above\_Median)) +  
 geom\_bar()  
pn34 = ggplot(ames\_salesClean, aes(x=House\_Style, fill = Above\_Median)) +  
 geom\_bar()  
pn35 = ggplot(ames\_salesClean, aes(x=Roof\_Style, fill = Above\_Median)) +  
 geom\_bar()  
pn36 = ggplot(ames\_salesClean, aes(x=Roof\_Matl, fill = Above\_Median)) +  
 geom\_bar()  
pn37 = ggplot(ames\_salesClean, aes(x=Exterior\_1st, fill = Above\_Median)) +  
 geom\_bar()  
pn38 = ggplot(ames\_salesClean, aes(x=Exterior\_2nd, fill = Above\_Median)) +  
 geom\_bar()  
pn39 = ggplot(ames\_salesClean, aes(x=Mas\_Vnr\_Area, fill = Above\_Median)) +  
 geom\_bar()  
pn40 = ggplot(ames\_salesClean, aes(x=Foundation, fill = Above\_Median)) +  
 geom\_bar()  
pn41 = ggplot(ames\_salesClean, aes(x=Heating, fill = Above\_Median)) +  
 geom\_bar()  
pn42 = ggplot(ames\_salesClean, aes(x=Central\_Air, fill = Above\_Median)) +  
 geom\_bar()  
pn43 = ggplot(ames\_salesClean, aes(x=Garage\_Type, fill = Above\_Median)) +   
 geom\_bar()  
pn44 = ggplot(ames\_salesClean, aes(x=Misc\_Feature, fill = Above\_Median)) +   
 geom\_bar()  
pn45 = ggplot(ames\_salesClean, aes(x=Sale\_Type, fill = Above\_Median)) +   
 geom\_bar()  
pn46 = ggplot(ames\_salesClean, aes(x=Sale\_Condition, fill = Above\_Median)) +   
 geom\_bar()  
  
  
# grid.arrange(pn24,pn25,pn26,pn27,pn28,ncol=2)  
# grid.arrange(pn29,pn30,pn31,pn32,ncol=2)  
# grid.arrange(pn33,pn34,pn35,pn36,ncol=2)  
# grid.arrange(pn37,pn38,pn39,pn40,ncol=2)  
# grid.arrange(pn41,pn42,pn43,pn44,ncol=2)  
# grid.arrange(pn45,pn46,ncol=2)

Just like the previous plots for the Discrete variables, many of these variables have a large portion of their observations sitting in one level. And of these that sit in this one level, the Above\_Median response is evenly distributed for the most part.For the charts that do not have most of their observations sitting in one level, I expect to see a clear “winner” between “Yes” and “No” (Above\_Median) in for each level(or the majority of the levels).

Will explore MS\_SubClass, Lot\_Config, Neighborhood, Exterior\_1st, EXterior\_2nd, and Garage\_Type further.

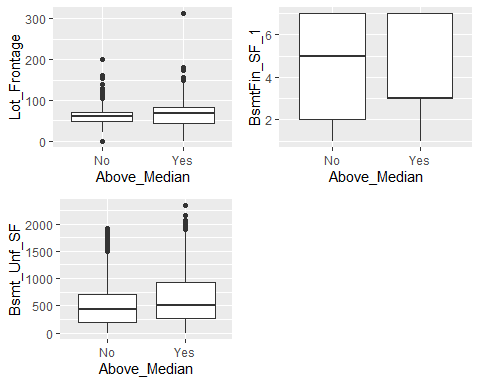
po1 = ggplot(ames\_salesClean, aes(x=Lot\_Shape, fill = Above\_Median)) +   
 geom\_bar()  
po2 = ggplot(ames\_salesClean, aes(x=Utilities, fill = Above\_Median)) +   
 geom\_bar()  
po3 = ggplot(ames\_salesClean, aes(x=Land\_Slope, fill = Above\_Median)) +   
 geom\_bar()  
po4 = ggplot(ames\_salesClean, aes(x=Overall\_Qual, fill = Above\_Median)) +   
 geom\_bar()+  
 xlab("Overall Quality")+  
 ylab("Count of Sales")+  
 labs(fill = "Above Median")+  
 theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))+  
 scale\_fill\_brewer(palette = "Set2")  
po5 = ggplot(ames\_salesClean, aes(x=Overall\_Cond, fill = Above\_Median)) +   
 geom\_bar()  
po6 = ggplot(ames\_salesClean, aes(x=Exter\_Qual, fill = Above\_Median)) +   
 geom\_bar()  
po7 = ggplot(ames\_salesClean, aes(x=Exter\_Cond, fill = Above\_Median)) +   
 geom\_bar()  
po8 = ggplot(ames\_salesClean, aes(x=Bsmt\_Qual, fill = Above\_Median)) +   
 geom\_bar()  
po9 = ggplot(ames\_salesClean, aes(x=Bsmt\_Cond, fill = Above\_Median)) +   
 geom\_bar()  
po10 = ggplot(ames\_salesClean, aes(x=Bsmt\_Exposure, fill = Above\_Median)) +   
 geom\_bar()  
po11 = ggplot(ames\_salesClean, aes(x=BsmtFin\_Type\_1, fill = Above\_Median)) +   
 geom\_bar()  
po12 = ggplot(ames\_salesClean, aes(x=BsmtFin\_Type\_2, fill = Above\_Median)) +   
 geom\_bar()  
po13 = ggplot(ames\_salesClean, aes(x=Heating\_QC, fill = Above\_Median)) +   
 geom\_bar()  
po14 = ggplot(ames\_salesClean, aes(x=Electrical, fill = Above\_Median)) +   
 geom\_bar()  
po15 = ggplot(ames\_salesClean, aes(x=Kitchen\_Qual, fill = Above\_Median)) +   
 geom\_bar()  
po16 = ggplot(ames\_salesClean, aes(x=Functional, fill = Above\_Median)) +   
 geom\_bar()+  
 xlab("Functional")+  
 ylab("Count of Sales")+  
 labs(fill = "Above Median")+  
 theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust=1))+  
 scale\_fill\_brewer(palette = "Set2")  
po17 = ggplot(ames\_salesClean, aes(x=Fireplace\_Qu, fill = Above\_Median)) +   
 geom\_bar()  
po18 = ggplot(ames\_salesClean, aes(x=Garage\_Finish, fill = Above\_Median)) +   
 geom\_bar()  
po19 = ggplot(ames\_salesClean, aes(x=Garage\_Qual, fill = Above\_Median)) +   
 geom\_bar()  
po20 = ggplot(ames\_salesClean, aes(x=Garage\_Cond, fill = Above\_Median)) +   
 geom\_bar()  
po21 = ggplot(ames\_salesClean, aes(x=Paved\_Drive, fill = Above\_Median)) +   
 geom\_bar()  
po22 = ggplot(ames\_salesClean, aes(x=Pool\_QC, fill = Above\_Median)) + geom\_bar()  
po23 = ggplot(ames\_salesClean, aes(x=Fence, fill = Above\_Median)) + geom\_bar()  
  
  
  
# grid.arrange(po1,po2,po3,po4,ncol=2)  
# grid.arrange(po5,po6,po7,po8,ncol=2)  
# grid.arrange(po9,po10,po11,po12,ncol=2)  
# grid.arrange(po13,po14,po15,po16,ncol=2)  
# grid.arrange(po17,po18,po19,po20,ncol=2)  
# grid.arrange(po21,po22,po23,ncol=2)

Just like the other variable types, there are variables that have a large portion of their observations sitting in one level. And of these that sit in this one level, the Above\_Median response is evenly distributed for the most part.

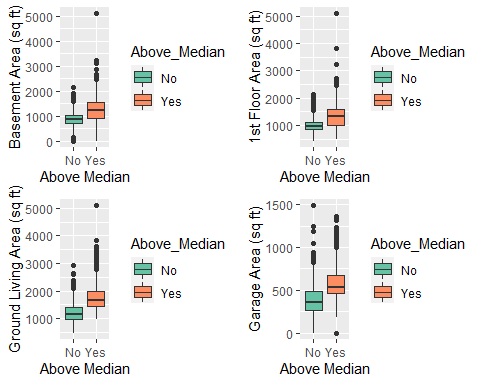
For the charts that do not have most of their observations sitting in one level, I expect to see a pattern e.g. “Yes” proportion of responses clearly increases or decreases from left to right since these are ordinal variables.

Will explore Lot\_Shape, Overall\_Qual, Overall\_Cond, Exter\_qual, Bsmt\_Qual, Bsmt\_Exposure\_, BsmtFin\_Type\_1, Heating\_QC, KItchen\_Qual, Fireplace\_Qu, Grage\_Finish further.

b1 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Lot\_Frontage))+  
 geom\_boxplot()  
b2 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=BsmtFin\_SF\_1))+  
 geom\_boxplot()  
b3 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Bsmt\_Unf\_SF))+  
 geom\_boxplot()  
b4 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Total\_Bsmt\_SF, fill=Above\_Median))+  
 geom\_boxplot()+  
 xlab("Above Median")+  
 ylab("Basement Area (sq ft)")+  
 scale\_fill\_brewer(palette = "Set2")  
b5 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=First\_Flr\_SF,   
 fill=Above\_Median))+  
 geom\_boxplot()+  
 xlab("Above Median")+  
 ylab("1st Floor Area (sq ft)")+  
 scale\_fill\_brewer(palette = "Set2")  
b6 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Gr\_Liv\_Area,  
 fill=Above\_Median))+  
 geom\_boxplot()+  
 xlab("Above Median")+  
 ylab("Ground Living Area (sq ft)")+  
 scale\_fill\_brewer(palette = "Set2")  
b7 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Garage\_Area, fill=Above\_Median))+  
 geom\_boxplot()+  
 xlab("Above Median")+  
 ylab("Garage Area (sq ft)")+  
 scale\_fill\_brewer(palette = "Set2")  
  
b8 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Latitude))+  
 geom\_boxplot()  
b9 = ggplot(ames\_salesClean, aes(x=Above\_Median, y=Longitude))+  
 geom\_boxplot()  
  
grid.arrange(b1,b2,b3,ncol=2)



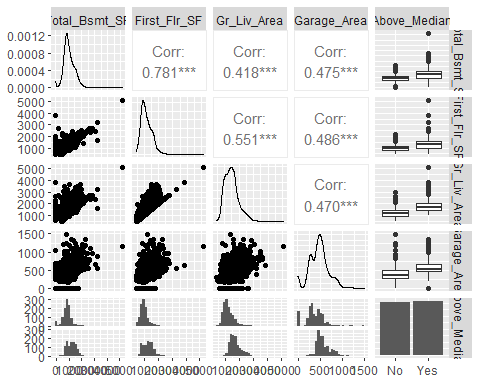
grid.arrange(b4,b5,b6,b7, ncol=2)

 Continuous variables that had a pretty nice looking histogram; Will explore Total\_Bsmt\_SF, First\_Flr\_SF, Gr\_LIv\_Area, and Garage\_Area a little further as scatter plots.

I am on the fence on exploring Latitude and Longitude further, since I believe these two will be closely related to the Neighborhood variable.

ggpairs(ames\_salesClean, columns = c(39, 44, 47, 62, 82))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



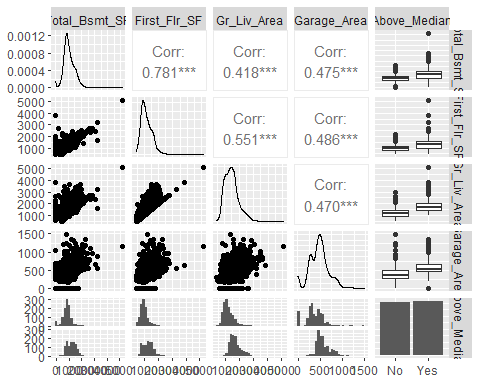
#Note: there is a 0.781 correlation between First\_Flr\_SF and Total\_Bsmt\_SF

Continuous variables that had a pretty nice looking histogram; Will explore Total\_Bsmt\_SF, First\_Flr\_SF, Gr\_LIv\_Area, and Garage\_Area a little further as scatter plots.

I am on the fence on exploring Latitude and Longitude further, since I believe these two will be closely related to the Neighborhood variable.

ggpairs(ames\_salesClean, columns = c(39, 44, 47, 62, 82))

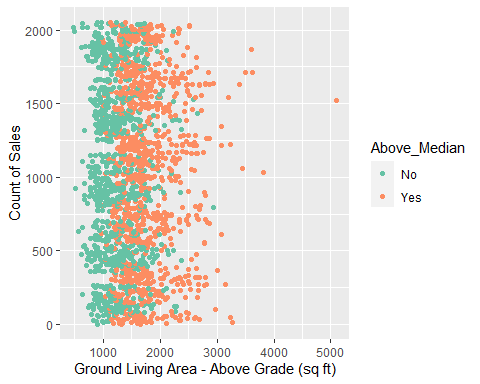
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



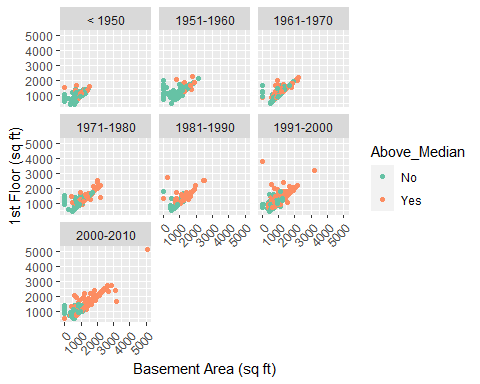
#Note: there is a 0.781 correlation between First\_Flr\_SF and Total\_Bsmt\_SF

Using ggpairs again to explore First\_Flr\_SF, Gr\_Liv\_Area, Garage\_Area, and Total\_Bsmt\_SF as predictors for Above\_Median.

ggplot(ames\_salesClean, aes(x=Gr\_Liv\_Area, y=X1,color=Above\_Median))+  
 geom\_point()+  
 xlab("Ground Living Area - Above Grade (sq ft)") + ylab("Count of Sales")+  
 labs(fill = "Above Median")+  
 scale\_color\_brewer(palette = "Set2")



# ggplot(ames\_salesClean, aes(x=First\_Flr\_SF, y=X1,  
# color=Above\_Median))+geom\_point()  
  
ggplot(ames\_salesClean, aes(x=Total\_Bsmt\_SF, y=First\_Flr\_SF, color = Above\_Median))+  
 geom\_point()+  
 xlab("Basement Area (sq ft)") + ylab("1st Floor (sq ft)")+  
 labs(fill = "Above Median")+  
 theme(axis.text.x = element\_text(angle = 46, vjust = 1, hjust=1))+  
 facet\_wrap(~ Year\_RemodAdd\_Group)+  
 scale\_color\_brewer(palette = "Set2")



# ggplot(ames\_salesClean, aes(x=Year\_Remod\_Add, y=First\_Flr\_SF, color = Above\_Median))+  
# geom\_point()+  
# theme(axis.text.x = element\_text(angle = 46, vjust = 1, hjust=1))+  
# facet\_wrap(~ Overall\_Qual)

Taking a deeper look into the 0.781 correlation between First Floor and Total Basement variables. Both of these variables, when created as Box Plots against Above\_Median, warrant a closer analysis

Although “Yes” observations tend to lean towards the higher BSmt SF and First Floor SF, there is still some overlapping, however, broken out by Year RemodAdd Group

p1 = ggplot(ames\_salesClean, aes(x=Lot\_Shape, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p2 = ggplot(ames\_salesClean, aes(x=Lot\_Config, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p3 = ggplot(ames\_salesClean, aes(x=Overall\_Qual, fill = Above\_Median)) +   
 geom\_bar(position="fill")+  
 xlab("Overall Quality") +  
 labs(fill = "Above Median")+  
 theme(axis.text.x = element\_text(angle = 46, vjust = 1, hjust=1))  
p4 = ggplot(ames\_salesClean, aes(x=Overall\_Cond, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p5 = ggplot(ames\_salesClean, aes(x=Bsmt\_Qual, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p6 = ggplot(ames\_salesClean, aes(x=Bsmt\_Exposure, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p7 = ggplot(ames\_salesClean, aes(x=BsmtFin\_Type\_1, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p8 = ggplot(ames\_salesClean, aes(x=Heating\_QC, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p9 = ggplot(ames\_salesClean, aes(x=Kitchen\_Qual, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p10 = ggplot(ames\_salesClean, aes(x=Garage\_Finish, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p11 = ggplot(ames\_salesClean, aes(x=Neighborhood, fill = Above\_Median)) +  
 geom\_bar(position="fill")+  
 ylab("Count of Sales")+  
 labs(fill = "Above Median")+  
 theme(axis.text.x = element\_text(angle = 90, vjust = 1, hjust=1))+  
 scale\_fill\_brewer(palette = "Set2")  
p12 = ggplot(ames\_salesClean, aes(x=Exterior\_1st, fill = Above\_Median)) +  
 geom\_bar(position="fill")  
p13 = ggplot(ames\_salesClean, aes(x=Exterior\_2nd, fill = Above\_Median)) +  
 geom\_bar(position="fill")  
p14 = ggplot(ames\_salesClean, aes(x=Bedroom\_AbvGr, fill = Above\_Median)) +   
 geom\_bar()+  
 ylab("Count of Sales")+  
 labs(fill = "Above Median")+  
 scale\_fill\_brewer(palette = "Set2")+  
 facet\_wrap(~ Year\_RemodAdd\_Group)  
p15 = ggplot(ames\_salesClean, aes(x=TotRms\_AbvGrd, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p16 = ggplot(ames\_salesClean, aes(x=Fireplaces, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p17 = ggplot(ames\_salesClean, aes(x=Fireplace\_Qu, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p18 = ggplot(ames\_salesClean, aes(x=Garage\_Type, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p19 = ggplot(ames\_salesClean, aes(x=Garage\_Cars, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p20 = ggplot(ames\_salesClean, aes(x=Exter\_Qual, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p21 = ggplot(ames\_salesClean, aes(x=Year\_Built\_Group, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p22 = ggplot(ames\_salesClean, aes(x=Year\_RemodAdd\_Group, fill = Above\_Median)) +  
 geom\_bar(position="fill")  
p23 = ggplot(ames\_salesClean, aes(x=Full\_Bath, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
p24 = ggplot(ames\_salesClean, aes(x=MS\_SubClass, fill = Above\_Median)) +   
 geom\_bar(position="fill")  
  
  
# grid.arrange(p1,p2,p3,p4,ncol=2)  
# grid.arrange(p5,p6,p7,p8,ncol=2)  
# grid.arrange(p9,p10,p11,p12,ncol=2)  
# grid.arrange(p13,p14,p15,p16,ncol=2)  
# grid.arrange(p17,p18,p19,p20,ncol=2)  
# grid.arrange(p21,p22,p23,p24,ncol=2)

Remove Lot\_Config, Exterior\_2nd, Bedroom\_AbvGr, Overall\_Cond, and BSmtFin\_Type1 from further analysis as strong predictor variables for Above\_Median.

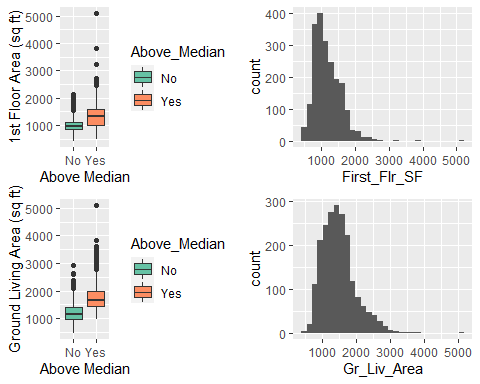
Viewing a filled bar graph for these variables make it a little more clear that the “Yes” and “No” variables are more evenly distributed across the levels, not good for predicting.

# ames\_salesClean %>% count(First\_Flr\_SF)  
# ames\_salesClean %>% count(Gr\_Liv\_Area)  
# ames\_salesClean %>% count(Garage\_Area)  
# ames\_salesClean %>% count(Total\_Bsmt\_SF)

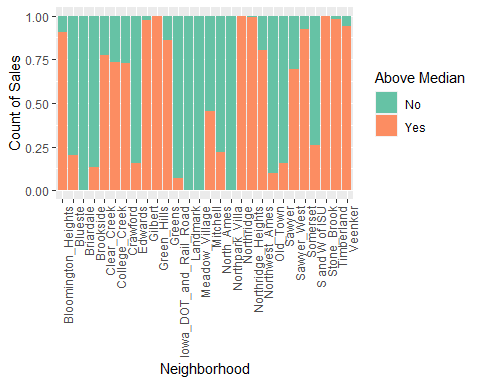
Removing Garage\_Area and Total\_Bsmt\_SF as strong predictors due to the 0 sqf observations.

grid.arrange(b5,h8,b6,h11,ncol=2)

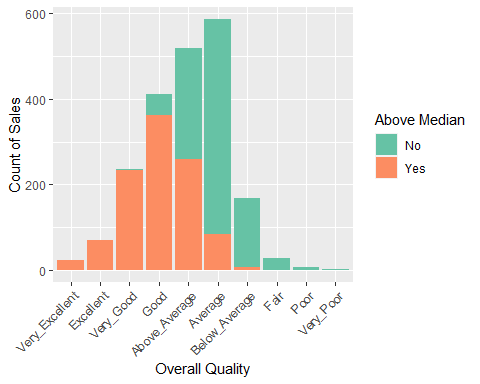
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



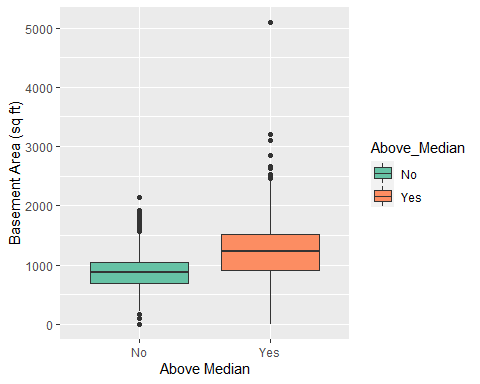
# p22  
# p15  
p11



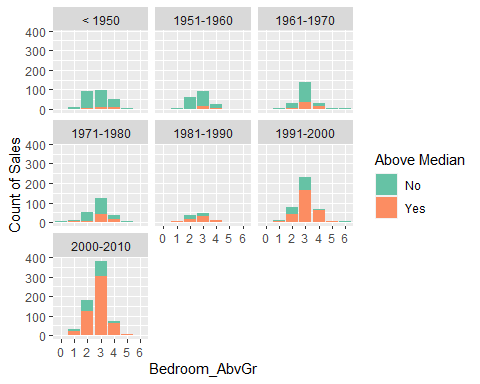
# p1  
po4



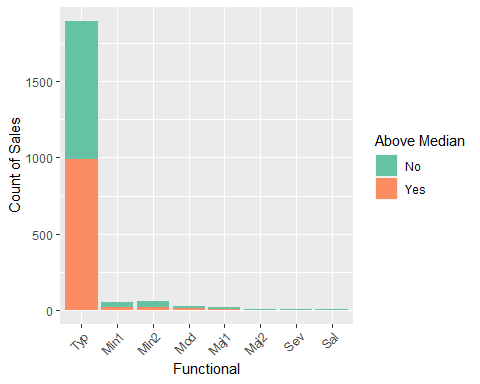
# p3  
# p20  
# p9  
# p10  
b4



p14



po16



# grid.arrange(b7,h12, b4,h7, ncol=2)  
#   
# p21  
# p16  
# p19  
# p23  
# p18  
# p5  
# p8  
# p17