# Final Project Phase 1 (EDA)

## Evan Hildreth

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

## Rows: 2053 Columns: 81

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (47): MS\_SubClass, MS\_Zoning, Street, Alley, Lot\_Shape, Land\_Contour, Ut...  
## dbl (34): Lot\_Frontage, Lot\_Area, Year\_Built, Year\_Remod\_Add, Mas\_Vnr\_Area, ...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#summary(ames)

#str(ames)

#skim(ames)

First I ran summary, stry, and skim to get an idea of the dataset that I am working with here. Also to check if there are any missing values that needed to be cleaned. Thankfully, this dataset is clean and there are not any NA’s as evidenced by the “skim” output.

ames\_select <- ames %>% mutate\_if(sapply(ames, is.character), as.factor)

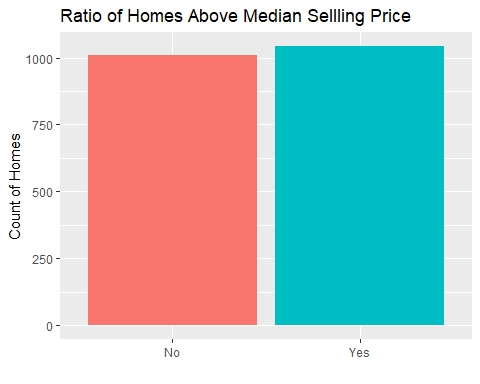
To begin the EDA I went ahead and converted all character (categories) to factors as they need to be factors to be compared to “Above Median.” In the future with our EDA.

ames\_select <- ames\_select %>% dplyr::select(-"Longitude", -"Latitude")

Next, I decided to cut out Longitude and Latitude from the analysis as the format of this quantitative variable is, by its very nature, so very different from every other quantitative variable in the dataset it could give some confusing results. Further, it is a bit too precise of a variable so each degree of Longitude and Latitude would likely only have one or two houses.

Since the response variable, Above\_Median, is the most important, it would be best to start the EDA by looking at the breakdown of this variable.

ggplot(ames\_select, aes(x=Above\_Median, fill= Above\_Median)) + geom\_bar() +  
 ggtitle("Ratio of Homes Above Median Sellling Price") + xlab("") +ylab("Count of Homes") + guides(fill="none")



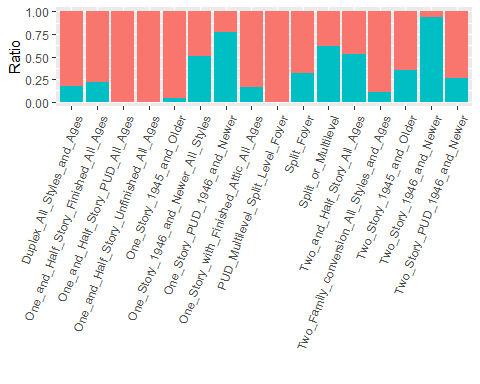
What we found is that there were slightly more houses sold above the median than bellow, which would indicate that we have some skew in this dataset.

### Categorial Variables

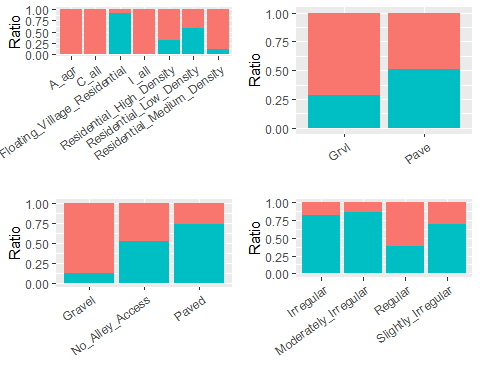
Next we are going to make a series of plot to compare the categorical (factor) data to help us determine what variables might be good to use when we later build our predictive model. The main plot tool of this will be to make bar graphs with using Above\_Median for the fill function. This method will be done to see if there are any variables that have a strong tenancy towards being above or bellow median. For sake of clarity it will be done in grids of 4 as that seems to be the ‘sweet spot’ for best use of space and readability.

In further analysis of these graphs, the most interesting and/or significant comparison plots will be presented to stakeholders.

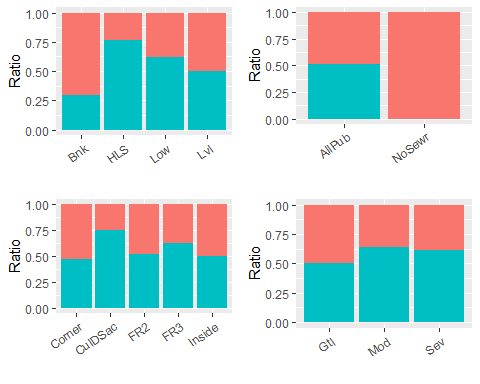
ggplot(ames\_select, aes(x = MS\_SubClass, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 70, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")



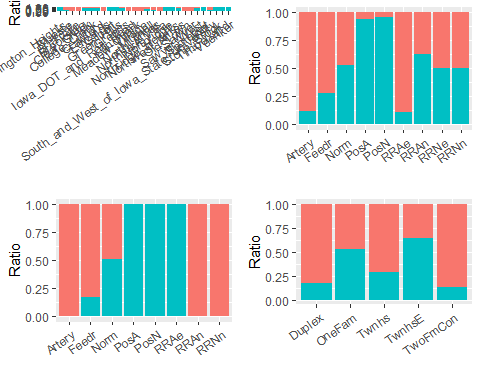
p2 <- ggplot(ames\_select, aes(x = MS\_Zoning, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p3 <- ggplot(ames\_select, aes(x = Street, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p4 <- ggplot(ames\_select, aes(x = Alley, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p5 <- ggplot(ames\_select, aes(x = Lot\_Shape, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
grid.arrange(p2,p3,p4,p5)

 From these first 5 plots it looks that Alley access has some pretty strong patterns and has only a few categories so it might be easier to work with.

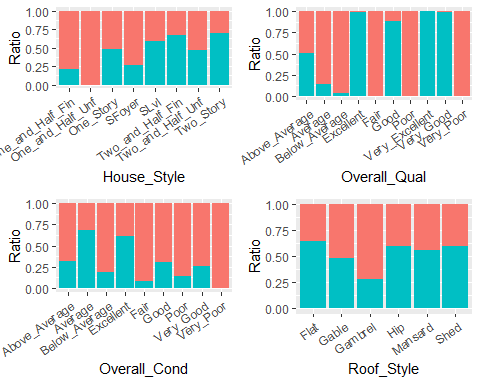
p6 <- ggplot(ames\_select, aes(x = Land\_Contour, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p7 <- ggplot(ames\_select, aes(x = Utilities, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p8 <- ggplot(ames\_select, aes(x = Lot\_Config, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p9 <- ggplot(ames\_select, aes(x = Land\_Slope, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
grid.arrange(p6,p7,p8,p9)

 From this next set of 4 it seems that Utilities is is pretty much a given for a home to be above\_median but still not the strongest indicator.

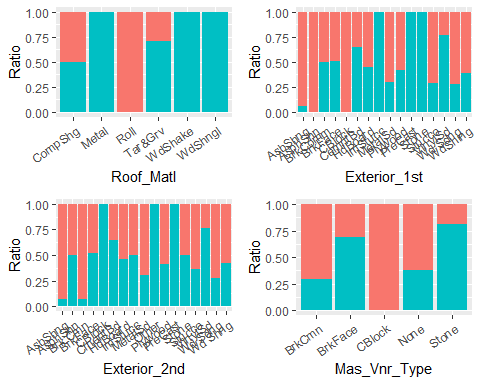
p10 <- ggplot(ames\_select, aes(x = Neighborhood, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p11 <- ggplot(ames\_select, aes(x = Condition\_1, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p12 <- ggplot(ames\_select, aes(x = Condition\_2, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
p13 <- ggplot(ames\_select, aes(x = Bldg\_Type, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") + guides(fill="none")  
grid.arrange(p10,p11,p12,p13)

 From these variables, it seems that neighborhood has the most variance with regards to the response variable in that some neighborhoods have a very high Above Median ratio while others are very low. Condition 1 and 2 might be worth investigation as well but may not as strong.

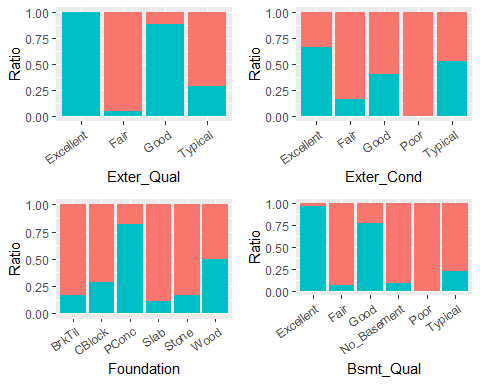
p14 <- ggplot(ames\_select, aes(x = House\_Style, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p15 <- ggplot(ames\_select, aes(x = Overall\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p16 <- ggplot(ames\_select, aes(x = Overall\_Cond, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p17 <- ggplot(ames\_select, aes(x = Roof\_Style, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
  
grid.arrange(p14,p15,p16,p17)

 From these I would say that House\_Style and Overall\_Qual seem to be the best indicators with House\_style having a very clear upgrades grade in above avg when the home gets bigger and the better the quality the higher the ratio of above median. Overall Condition might be useful as a backup but Roof\_style seems to have a fairly small impact.

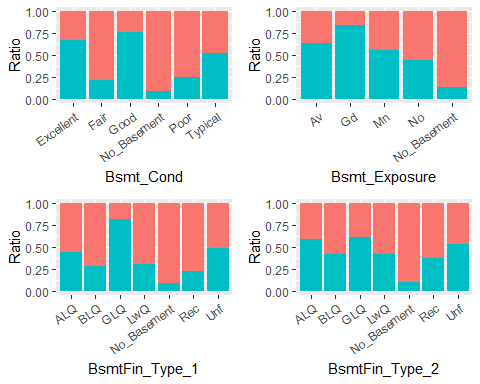
p18 <- ggplot(ames\_select, aes(x = Roof\_Matl, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p19 <- ggplot(ames\_select, aes(x = Exterior\_1st, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p20 <- ggplot(ames\_select, aes(x = Exterior\_2nd, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p21 <- ggplot(ames\_select, aes(x = Mas\_Vnr\_Type, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p18,p19,p20,p21)

 Mas\_Vnr\_Type seems to be a potential good indicator as Stone and BrkFace has a very good chance of being Above\_Median. Roof\_Matl also may be good as any roof made of Roll is seemingly guaranteed to be bellow median but the other material types are either 100% or roughly 50/50.

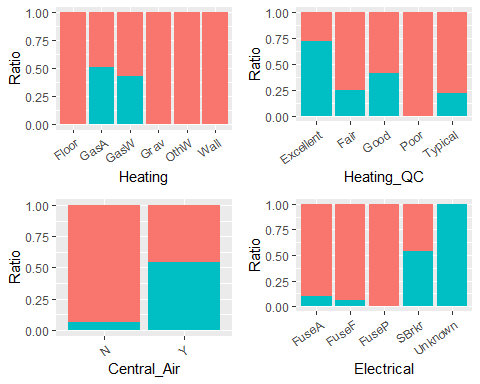
p22 <- ggplot(ames\_select, aes(x = Exter\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p23 <- ggplot(ames\_select, aes(x = Exter\_Cond, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p24 <- ggplot(ames\_select, aes(x = Foundation, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p25 <- ggplot(ames\_select, aes(x = Bsmt\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p22,p23,p24,p25)

 The two Qual variables seem to best pretty strong indicators as if the exterior or the basement is of “Excellent” or “Good” quality it has a very high chance of being Above\_Median.

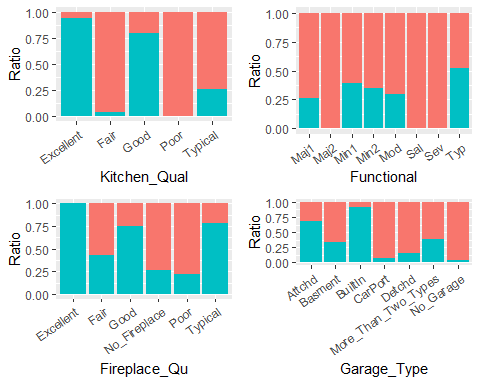
p26 <- ggplot(ames\_select, aes(x = Bsmt\_Cond, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p27 <- ggplot(ames\_select, aes(x = Bsmt\_Exposure, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p28 <- ggplot(ames\_select, aes(x = BsmtFin\_Type\_1, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p29 <- ggplot(ames\_select, aes(x = BsmtFin\_Type\_2, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p26,p27,p28,p29)

 From all 4 of these variables not having a basement is a pretty major determent to the value of the home, but beyond that there does not seem to be any major indicators which would lead to an above median home value.

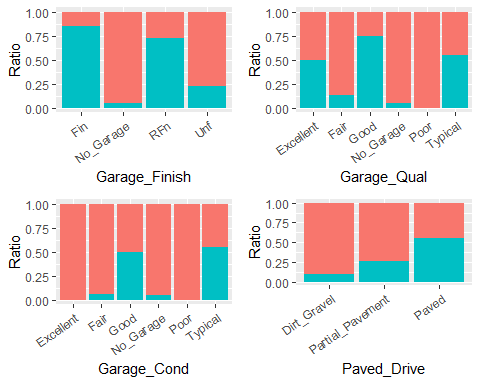
p30 <- ggplot(ames\_select, aes(x = Heating, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p31 <- ggplot(ames\_select, aes(x = Heating\_QC, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p32 <- ggplot(ames\_select, aes(x = Central\_Air, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p33 <- ggplot(ames\_select, aes(x = Electrical, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p30,p31,p32,p33)

 Not particularly a fan of any of these variables. Electrical defiantly should not be used with the major “Unknown” category. The only ones is that there are some heating types that just don’t apply to Above median houses.

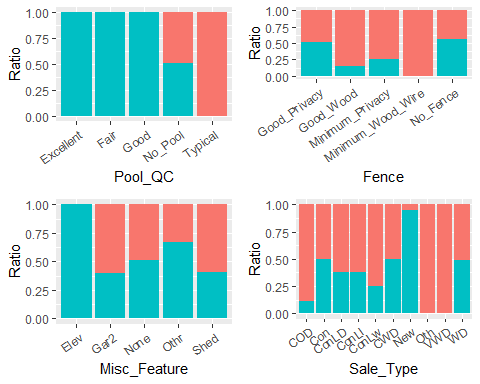
p34 <- ggplot(ames\_select, aes(x = Kitchen\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p35 <- ggplot(ames\_select, aes(x = Functional, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p36 <- ggplot(ames\_select, aes(x = Fireplace\_Qu, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p37 <- ggplot(ames\_select, aes(x = Garage\_Type, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p34,p35,p36,p37)

 There are some good variables here to work with. Kitchen\_Qual an Fireplace\_Qu seem to highly favor excellent and good (lesser extend typical) for Above\_median homes. Built In garages also looks to be a very strong indicator of being Above\_median with no\_garage being the opposite.

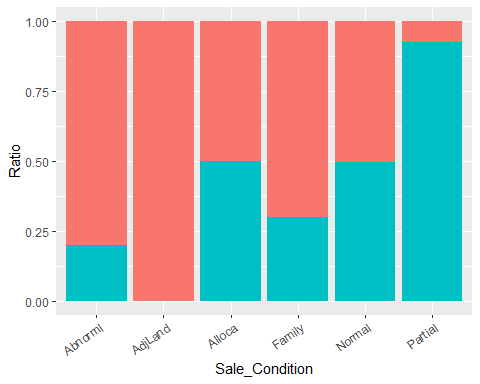
p38 <- ggplot(ames\_select, aes(x = Garage\_Finish, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p39 <- ggplot(ames\_select, aes(x = Garage\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p40 <- ggplot(ames\_select, aes(x = Garage\_Cond, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p41 <- ggplot(ames\_select, aes(x = Paved\_Drive, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p38,p39,p40,p41)

 The best variable from this bunch seems to be Paved\_drive as there is a pretty clear distinction between the 3 types. Surprisingly, garage quality seems to be not have as strong of an indication compared to other quality variables tested, except that having not garage or no garage is almost always with bellow median homes.

p42 <- ggplot(ames\_select, aes(x = Pool\_QC, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p43 <- ggplot(ames\_select, aes(x = Fence, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p44 <- ggplot(ames\_select, aes(x = Misc\_Feature, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p45 <- ggplot(ames\_select, aes(x = Sale\_Type, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
grid.arrange(p42,p43,p44,p45)

 From these list of variables, the most interesting is the Pool\_QC in that practically every home that has a pool was sold above median (with the lone exception being the ‘typical’ rating). This seems like a good variable to use in that if a home has a pool there is a good change it was sold Above Median. Fence does not seem to have a major impact, but Misc\_Feature could in that Bev is almost associated with Above median.

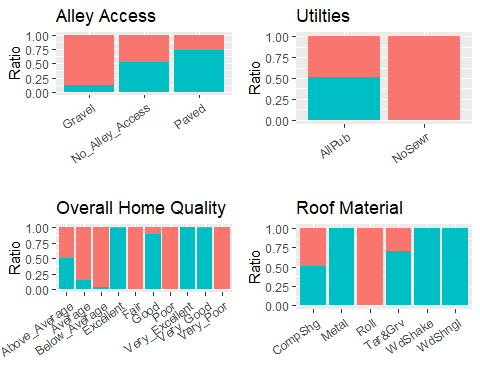
p46 <- ggplot(ames\_select, aes(x = Sale\_Condition, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + guides(fill="none")  
p46

 For the last categorical variable it looks as if this might be a good one as partial seems to be a strong indicator of being above median.

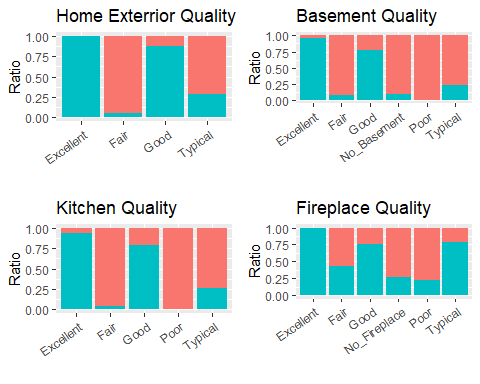
### Categorial Best Variables

Bellow the line will be what I consider to be the 13 most interesting and useful categorical variables that I will likely be using in future state for my prediction analysis of “Above\_Median” The general trends found here is that having better amenities, living in better neighborhoods, and having higher quality homes (garage, interior, exterior etc) generally leads to homes begin sold above median.

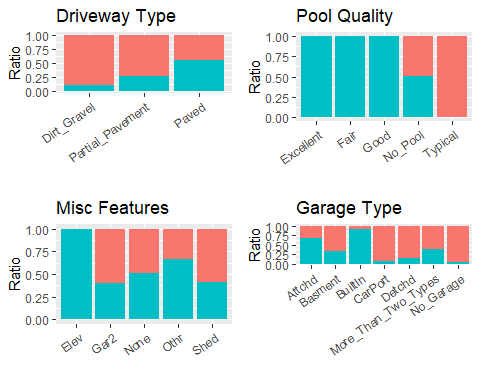
p4 <- ggplot(ames\_select, aes(x = Alley, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Alley Access") + xlab("") + guides(fill="none")  
p7 <- ggplot(ames\_select, aes(x = Utilities, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + ggtitle("Utilties") +xlab("") + guides(fill="none")  
p15 <- ggplot(ames\_select, aes(x = Overall\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + ggtitle("Overall Home Quality") +xlab("") + guides(fill="none")  
p18 <- ggplot(ames\_select, aes(x = Roof\_Matl, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") + ggtitle("Roof Material") +xlab("") + guides(fill="none")  
grid.arrange(p4,p7,p15,p18)



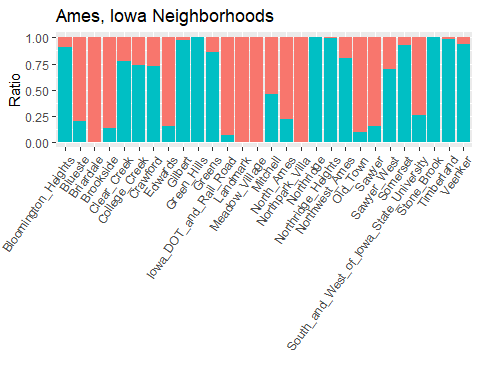
p22 <- ggplot(ames\_select, aes(x = Exter\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Home Exterrior Quality") + xlab("") + guides(fill="none")  
p25 <- ggplot(ames\_select, aes(x = Bsmt\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Basement Quality") + xlab("") + guides(fill="none")  
p34 <- ggplot(ames\_select, aes(x = Kitchen\_Qual, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Kitchen Quality") + xlab("") + guides(fill="none")  
p36 <- ggplot(ames\_select, aes(x = Fireplace\_Qu, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Fireplace Quality") + xlab("") + guides(fill="none")  
grid.arrange(p22,p25,p34,p36)



p41 <- ggplot(ames\_select, aes(x = Paved\_Drive, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Driveway Type") + xlab("") + guides(fill="none")  
p42 <- ggplot(ames\_select, aes(x = Pool\_QC, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Pool Quality") + xlab("") + guides(fill="none")  
p44 <- ggplot(ames\_select, aes(x = Misc\_Feature, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Misc Features") + xlab("") + guides(fill="none")  
p37 <- ggplot(ames\_select, aes(x = Garage\_Type, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio") +ggtitle("Garage Type") + xlab("") + guides(fill="none")  
grid.arrange(p41,p42,p44,p37)



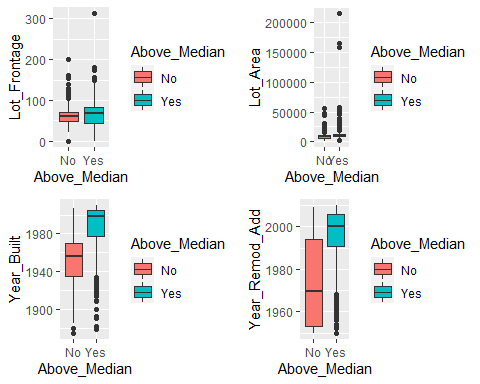
p10 <- ggplot(ames\_select, aes(x = Neighborhood, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 55, vjust = 1, hjust=1)) + ylab("Ratio") +xlab("") +ggtitle("Ames, Iowa Neighborhoods") + guides(fill="none")  
p10

 The above 13 plots seems to show the strongest quantitative indicators if a home was sold for above median pricing or not.

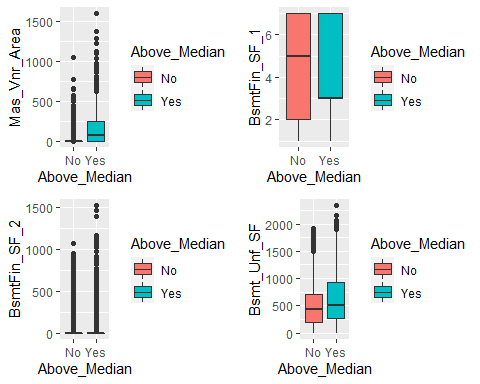
### Quantative Values

Next, we are going to move on to the next section of analysis which will be conducting an EDA on our Quantitative values. For this we are taking a different approach and are looking at box plots of the values broken down into “Yes” and “No” grouping. Again, for the sake of space conversation and readability the plots will be put in groupings of 4.

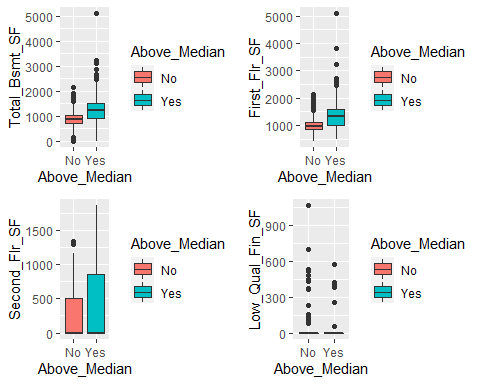
p47 <- ggplot(ames\_select, aes(x = Above\_Median, y = Lot\_Frontage, fill =Above\_Median)) + geom\_boxplot()  
p48 <- ggplot(ames\_select, aes(x = Above\_Median, y = Lot\_Area, fill =Above\_Median)) + geom\_boxplot()  
p49 <- ggplot(ames\_select, aes(x = Above\_Median, y = Year\_Built, fill =Above\_Median )) + geom\_boxplot()  
p50 <- ggplot(ames\_select, aes(x = Above\_Median, y = Year\_Remod\_Add, fill =Above\_Median)) +geom\_boxplot()  
grid.arrange(p47,p48,p49,p50)

 For these first for quantitative variables it seems pretty clear that the more recent the home was built or remodeled the more likely it is that the home was sold Above Median. Naturally, there are some exception and “Yes” tends to have a wider range but this is useful.

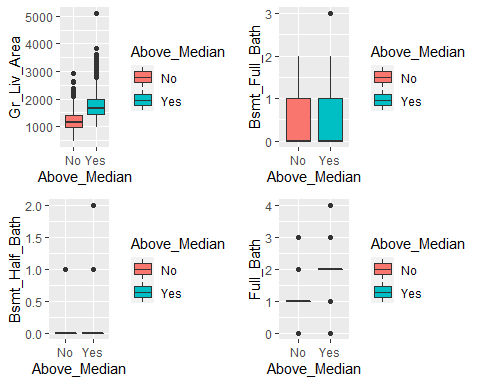
p51 <- ggplot(ames\_select, aes(x = Above\_Median, y = Mas\_Vnr\_Area, fill= Above\_Median)) + geom\_boxplot()  
p52 <- ggplot(ames\_select, aes(x = Above\_Median, y = BsmtFin\_SF\_1, fill= Above\_Median)) + geom\_boxplot()  
p53 <- ggplot(ames\_select, aes(x = Above\_Median, y = BsmtFin\_SF\_2, fill= Above\_Median)) + geom\_boxplot()  
p54 <- ggplot(ames\_select, aes(x = Above\_Median, y = Bsmt\_Unf\_SF, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p51,p52,p53,p54)

 Not really seeing anything that jumps out to me from these variables. Possibly Mas\_Vnr\_Area but that’s only if we need to find another one to work with.

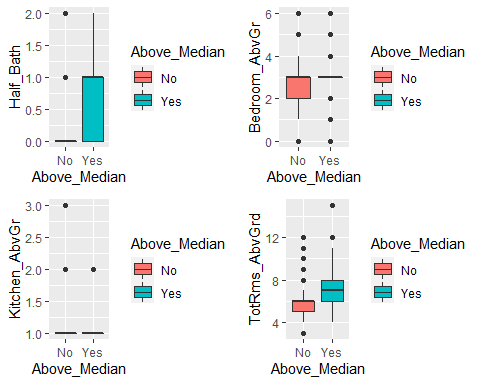
p55 <- ggplot(ames\_select, aes(x = Above\_Median, y = Total\_Bsmt\_SF, fill= Above\_Median)) + geom\_boxplot()  
p56 <- ggplot(ames\_select, aes(x = Above\_Median, y = First\_Flr\_SF, fill= Above\_Median)) + geom\_boxplot()  
p57 <- ggplot(ames\_select, aes(x = Above\_Median, y = Second\_Flr\_SF, fill= Above\_Median)) + geom\_boxplot()  
p58 <- ggplot(ames\_select, aes(x = Above\_Median, y = Low\_Qual\_Fin\_SF, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p55,p56,p57,p58)

 Looking at these variables in regards to square feet of different portions of the home, it does show that homes that sell Above Median typically have larger square feet than homes that sell bellow median. Maybe not the strongest indicator but good to hold onto if need be.

p59 <- ggplot(ames\_select, aes(x = Above\_Median, y = Gr\_Liv\_Area, fill= Above\_Median)) + geom\_boxplot()  
p60 <- ggplot(ames\_select, aes(x = Above\_Median, y = Bsmt\_Full\_Bath, fill= Above\_Median)) + geom\_boxplot()   
p61 <- ggplot(ames\_select, aes(x = Above\_Median, y = Bsmt\_Half\_Bath, fill= Above\_Median)) + geom\_boxplot()  
p62 <- ggplot(ames\_select, aes(x = Above\_Median, y = Full\_Bath, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p59,p60,p61,p62)

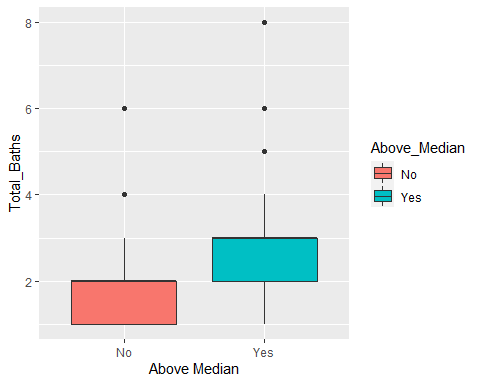
 Looking at these variables in regard to the last of the rooms SF it is on trend that the larger the living space the more likely the home is to sell above Median. Looking at the start of the bathroom count variables there does not seem to be any clear major trends. There seems to be one that Above median homes have more bathrooms but is is not super clear. It may be best to create a new variable in which all the bathrooms are added together,

p63 <- ggplot(ames\_select, aes(x = Above\_Median, y = Half\_Bath, fill= Above\_Median)) + geom\_boxplot()  
p64 <- ggplot(ames\_select, aes(x = Above\_Median, y = Bedroom\_AbvGr, fill= Above\_Median)) + geom\_boxplot()  
p65 <- ggplot(ames\_select, aes(x = Above\_Median, y = Kitchen\_AbvGr, fill= Above\_Median)) + geom\_boxplot()  
p66 <- ggplot(ames\_select, aes(x = Above\_Median, y = TotRms\_AbvGrd, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p63,p64,p65,p66)

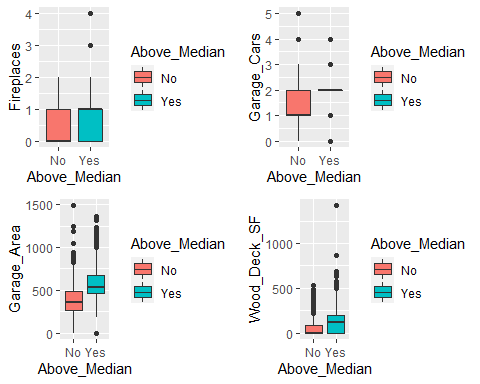
 Here it does seem to look that there is a trend that more bathrooms and total rooms are associated with Above Median homes.

ames\_select <- mutate(ames\_select, Total\_Baths = Half\_Bath + Bsmt\_Full\_Bath + Bsmt\_Half\_Bath + Full\_Bath)

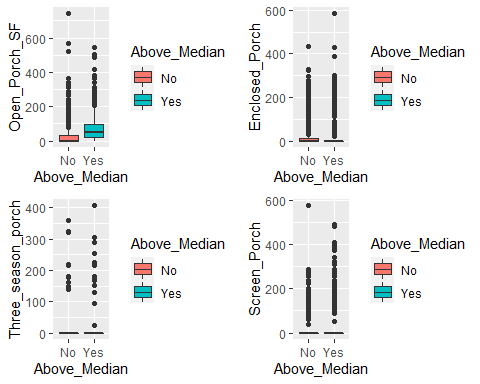
p67 <- ggplot(ames\_select, aes(x = Above\_Median, y = Total\_Baths, fill= Above\_Median)) + geom\_boxplot() +xlab("Above Median")  
p67

 Deciding to take a brief sidestep to create a new variable that adds all the baths together in a home and this paints a more clear picture that Above Median homes tend to have more bathrooms for any kind.

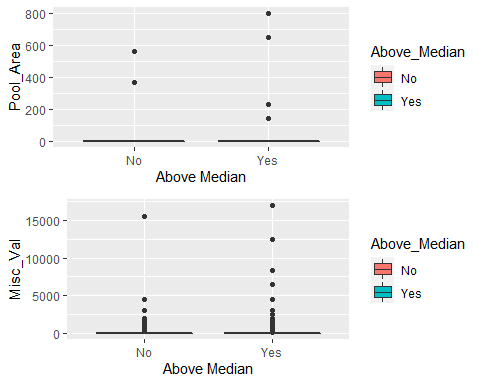
p68 <- ggplot(ames\_select, aes(x = Above\_Median, y = Fireplaces, fill= Above\_Median)) + geom\_boxplot()  
p69 <- ggplot(ames\_select, aes(x = Above\_Median, y = Garage\_Cars, fill= Above\_Median)) + geom\_boxplot()  
p70 <- ggplot(ames\_select, aes(x = Above\_Median, y = Garage\_Area, fill= Above\_Median)) + geom\_boxplot()  
p71 <- ggplot(ames\_select, aes(x = Above\_Median, y = Wood\_Deck\_SF, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p68,p69,p70,p71)

 Looking at these it again becomes clear that more square feet is associated with Above Median homes this is true for Garage area and wood deck but I think wood deck might not be the best to use because there were many zeros in the No category.

p72 <- ggplot(ames\_select, aes(x = Above\_Median, y = Open\_Porch\_SF, fill= Above\_Median)) + geom\_boxplot()  
p73 <- ggplot(ames\_select, aes(x = Above\_Median, y = Enclosed\_Porch, fill= Above\_Median)) + geom\_boxplot()  
p74 <- ggplot(ames\_select, aes(x = Above\_Median, y = Three\_season\_porch, fill= Above\_Median)) + geom\_boxplot()  
p75 <- ggplot(ames\_select, aes(x = Above\_Median, y = Screen\_Porch, fill= Above\_Median)) + geom\_boxplot()  
grid.arrange(p72,p73,p74,p75)

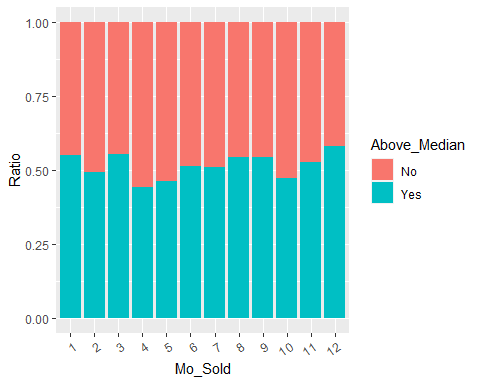
 Same General trend that more SF is more associated with Above Median Homes but not as strongly associated with these as many homes do not have these features, with the exception of Open porch which seems to be a stronger variable.

p76 <- ggplot(ames\_select, aes(x = Above\_Median, y = Pool\_Area, fill= Above\_Median)) + geom\_boxplot() +xlab("Above Median")  
p77 <- ggplot(ames\_select, aes(x = Above\_Median, y = Misc\_Val, fill = Above\_Median)) + geom\_boxplot() +xlab("Above Median")  
grid.arrange(p76,p77)

 Again, nothing really all that interesting to look at and no real strong trends to look at.

ames\_select <- ames\_select %>% mutate(Mo\_Sold = as\_factor(Mo\_Sold))

p78 <- ggplot(ames\_select, aes(x = Mo\_Sold, fill= Above\_Median)) +geom\_bar(position = "fill") + theme(axis.text.x = element\_text(angle = 35, vjust = 1, hjust=1)) + ylab("Ratio")  
p78

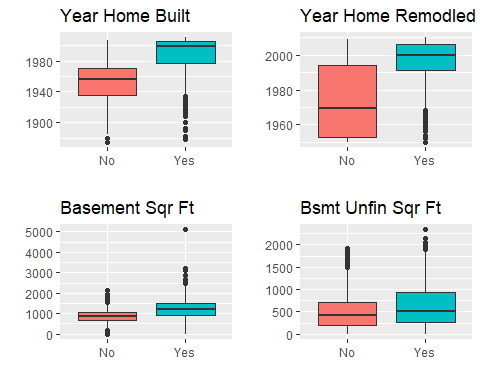
 For the last quantitative variable, months sold, it made more sense to treat it as a factor as it showed how likely a home was to be sold Above Median based off of what month it was sold in. From this last plot above we see that there does not really seem to be any major trends in this regard. There may be some potential seasonality but that’s a bit of a stretch.

### Quantative Values Result

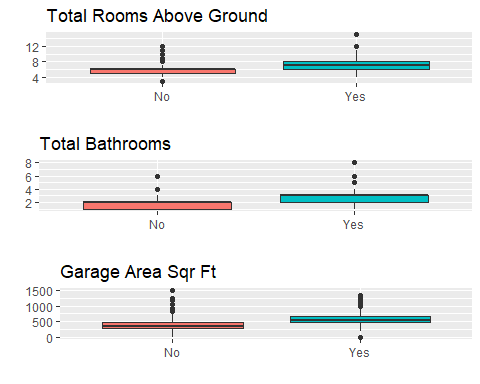
Compared to their qualitative / categorical values there simply was not as many variables to look at and on top of that there were even fewer that seemed to be candidates to be strong variables. However, bellow is the grouping of the most useful ones that I found.

The general trends to be found from this is that strong indicators that a home will be sold for above median pricing is that how recent the home was built/remodeled, square feet the place / each room has, and how many rooms (general and bathroom) with more almost always being better.

p49 <- ggplot(ames\_select, aes(x = Above\_Median, y = Year\_Built, fill =Above\_Median )) + geom\_boxplot() +xlab("") + ggtitle("Year Home Built") +ylab("") + guides(fill="none")  
p50 <- ggplot(ames\_select, aes(x = Above\_Median, y = Year\_Remod\_Add, fill =Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Year Home Remodled") +ylab("") + guides(fill="none")  
p54 <- ggplot(ames\_select, aes(x = Above\_Median, y = Bsmt\_Unf\_SF, fill= Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Bsmt Unfin Sqr Ft") +ylab("") + guides(fill="none")  
p55 <- ggplot(ames\_select, aes(x = Above\_Median, y = Total\_Bsmt\_SF, fill = Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Basement Sqr Ft") +ylab("") + guides(fill="none")  
grid.arrange(p49,p50,p55,p54)

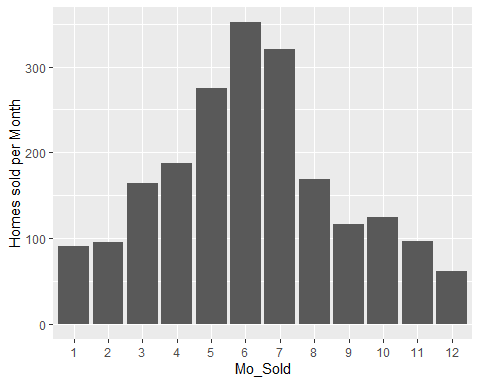


p66 <- ggplot(ames\_select, aes(x = Above\_Median, y = TotRms\_AbvGrd, fill= Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Total Rooms Above Ground") +ylab("") + guides(fill="none")  
p67 <- ggplot(ames\_select, aes(x = Above\_Median, y = Total\_Baths, fill = Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Total Bathrooms") +ylab("") + guides(fill="none")  
p70 <- ggplot(ames\_select, aes(x = Above\_Median, y = Garage\_Area, fill = Above\_Median)) + geom\_boxplot() +xlab("") + ggtitle("Garage Area Sqr Ft") +ylab("") + guides(fill="none")  
grid.arrange(p66,p67,p70)

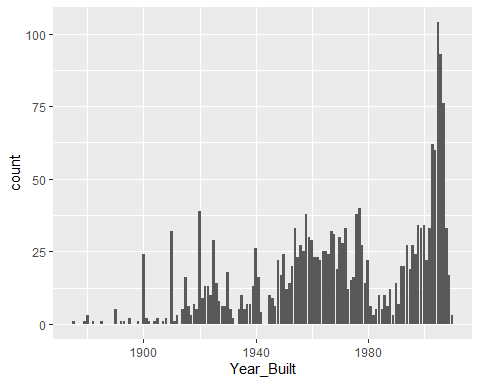
 ### Misc plots

While all the above plots serve their need to compare different variables with Above\_median, there are a few variables that deserve their own analysis and plots on their own to more fully understand the data. Namely this is to look at the distribution of homes sold by month, year built and remodeled over time and see how many homes are sold in each neighborhood.

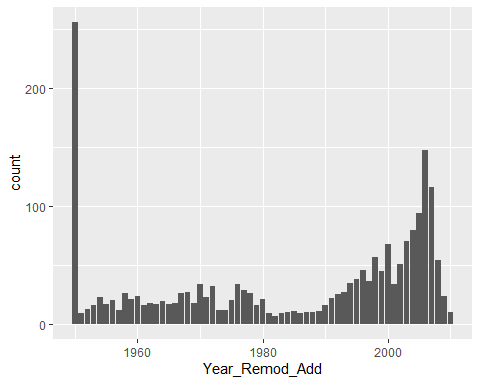
ggplot(ames\_select, aes(x=Mo\_Sold)) +geom\_bar() +ylab("Homes sold per Month")



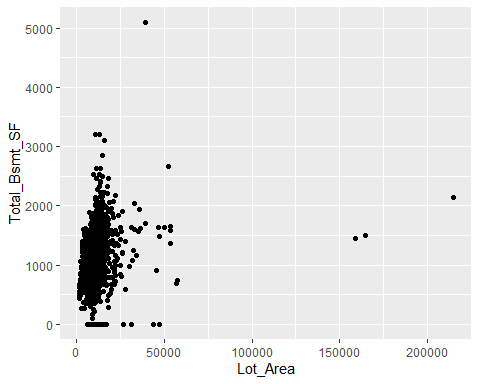
ggplot(ames\_select, aes(x=Year\_Built)) +geom\_bar()

 So far, we can see that most of the homes are sold in may, June, and July (5,6,7) an that most of the homes have been built within the past few years.

ggplot(ames\_select, aes(x=Year\_Remod\_Add)) +geom\_bar()

 Similar trend for home remodeling but with the major caveat that many of the homes list the remodel as 1950. This potentially could be due to the fact that they may have been older than 1950 but that’s as far back as the dataset goes.

ggplot(ames\_select, aes(x=Lot\_Area, y=Total\_Bsmt\_SF)) +geom\_point()

 This plot is to see if there is any clear relationship between the total lot size and the size of one of the rooms, namely the basement. If looking at the main cluster and ignoring the handful out outliers we do see a trend that as lot size increases total basement sf increases as well.