# BAN 502 Course Project

## Dawadi, Babu

### Phase 1 Project Description

Load Libraries:

#install.packages("rlang::last\_error()")  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidymodels)

## Warning: package 'tidymodels' was built under R version 4.1.2

## Registered S3 method overwritten by 'tune':  
## method from   
## required\_pkgs.model\_spec parsnip

## -- Attaching packages -------------------------------------- tidymodels 0.1.4 --

## v broom 0.7.9 v rsample 0.1.1   
## v dials 0.0.10 v tune 0.1.6   
## v infer 1.0.0 v workflows 0.2.4   
## v modeldata 0.1.1 v workflowsets 0.1.0   
## v parsnip 0.1.7 v yardstick 0.0.9   
## v recipes 0.1.17

## Warning: package 'dials' was built under R version 4.1.2

## Warning: package 'scales' was built under R version 4.1.2

## Warning: package 'infer' was built under R version 4.1.2

## Warning: package 'modeldata' was built under R version 4.1.2

## Warning: package 'parsnip' was built under R version 4.1.2

## Warning: package 'recipes' was built under R version 4.1.2

## Warning: package 'rsample' was built under R version 4.1.2

## Warning: package 'tune' was built under R version 4.1.2

## Warning: package 'workflows' was built under R version 4.1.2

## Warning: package 'workflowsets' was built under R version 4.1.2

## Warning: package 'yardstick' was built under R version 4.1.2

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## \* Dig deeper into tidy modeling with R at https://www.tmwr.org

library(GGally) #ggcorr and ggpairs

## Warning: package 'GGally' was built under R version 4.1.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(ggcorrplot) #correlation plot alternative

## Warning: package 'ggcorrplot' was built under R version 4.1.2

library(gridExtra) #create grids of plots

## Warning: package 'gridExtra' was built under R version 4.1.2

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(lubridate)

## Warning: package 'lubridate' was built under R version 4.1.2

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(car) #for the VIF function

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.1.2

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-3

### Task 1:

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)

Converting Character and Numeric variables to Factor as appropriate

# converting all character variables to factor  
ames = ames %>% mutate\_if(is.character,as\_factor)   
  
# converting hr from numeric to factor  
ames$BsmtFin\_SF\_1 <- as\_factor(ames$BsmtFin\_SF\_1)  
ames$Bsmt\_Full\_Bath <- as\_factor(ames$Bsmt\_Full\_Bath)  
ames$Bsmt\_Half\_Bath <- as\_factor(ames$Bsmt\_Half\_Bath)  
ames$Full\_Bath <- as\_factor(ames$Full\_Bath)  
ames$Half\_Bath <- as\_factor(ames$Half\_Bath)  
ames$Bedroom\_AbvGr <- as\_factor(ames$Bedroom\_AbvGr)  
ames$Kitchen\_AbvGr <- as\_factor(ames$Kitchen\_AbvGr)  
#ames$Low\_Qual\_Fin\_SF <- as.factor(ames$Low\_Qual\_Fin\_SF)  
  
ames$TotRms\_AbvGrd <- as\_factor(ames$TotRms\_AbvGrd)  
ames$Fireplaces <- as\_factor(ames$Fireplaces)  
ames$Garage\_Cars <- as\_factor(ames$Garage\_Cars)  
ames$Pool\_Area <- as\_factor(ames$Pool\_Area)  
  
ames = ames %>% mutate( Year\_Sold=as.Date(as.character(Year\_Sold), format = "%Y"))  
ames = ames %>% mutate( Mo\_Sold=as.Date(as.character(Mo\_Sold), format = "%Y"))

There are 2930 observations and 82 variables in Ames dataset and almost every variable assessed have contributed certain role in the housing prices. In some cases, the buyers may have their preference what type of home they are willing to purchase. Some variables have more impact on housing prices than the others.

After examining all the variables that affect the above median price (yes/no) the following 20 variables are considered to include in the analysis here. Here are those variables-

Numeric variables:

Original construction year

Remodeling year

Total square feet of basement area

First Floor square feet

Above grade (ground) living area square feet

Size of garage in square feet

Zoning classification

Type of dwelling involved

Neighborhood

Overall material and finish of the house

Categorical variables:

Exterior covering on house

External Quality of the house

Basement Quality

Heating Quality

Number of full bathrooms

Kitchen Quality

Total room above ground

Number of fireplaces

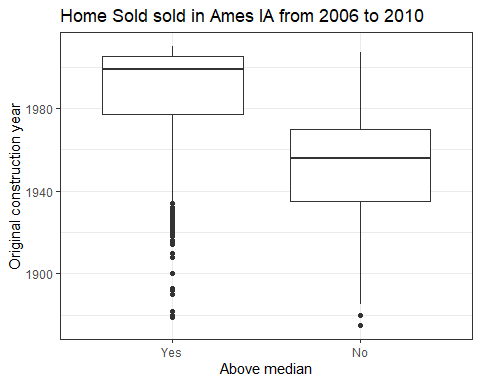
Garage type (location)

Size of garage in car capacity

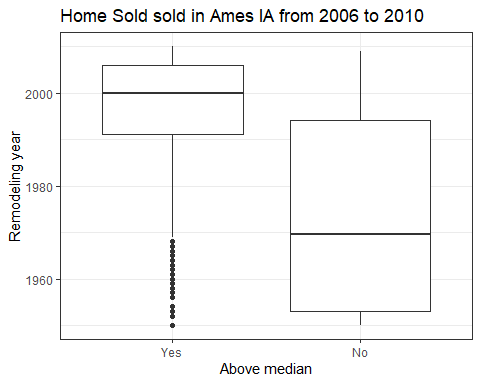
Let’s check the affect of the potential variables that on housing prices above median (yes/no) using visualization.

Box Plots for Numeric Variables

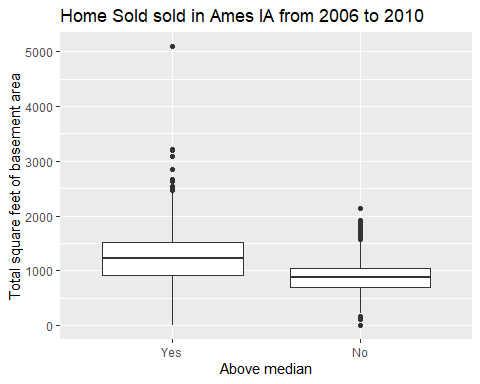
ggplot(ames,aes(x=Above\_Median,y=Year\_Built)) + geom\_boxplot() + theme\_bw() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "Original construction year")



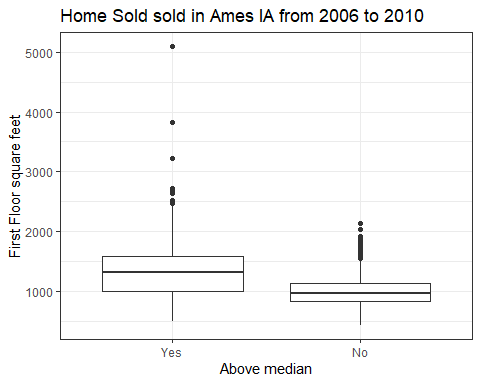
ggplot(ames,aes(x=Above\_Median,y=Year\_Remod\_Add)) + geom\_boxplot() + theme\_bw() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "Remodeling year")



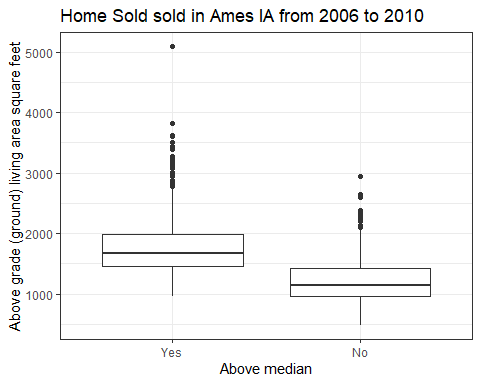
ggplot(ames,aes(x=Above\_Median,y=Total\_Bsmt\_SF)) + geom\_boxplot() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "Total square feet of basement area")



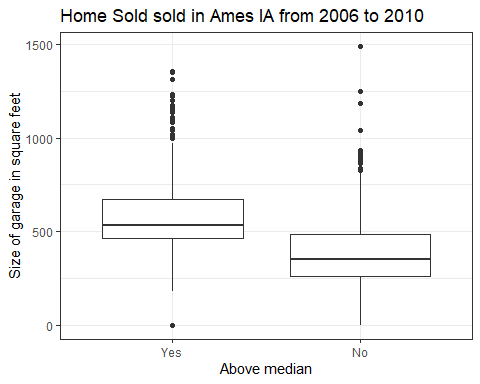
ggplot(ames,aes(x=Above\_Median,y=First\_Flr\_SF)) + geom\_boxplot() + theme\_bw() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "First Floor square feet")



ggplot(ames,aes(x=Above\_Median,y=Gr\_Liv\_Area)) + geom\_boxplot() + theme\_bw() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "Above grade (ground) living area square feet")



ggplot(ames,aes(x=Above\_Median,y=Garage\_Area)) + geom\_boxplot() + theme\_bw() +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Above median",  
 y = "Size of garage in square feet")



Here are the findings from box plots:

Newer houses (about 1970 and newer) have above median price.

Recent year remodeled house tend to have above median price.

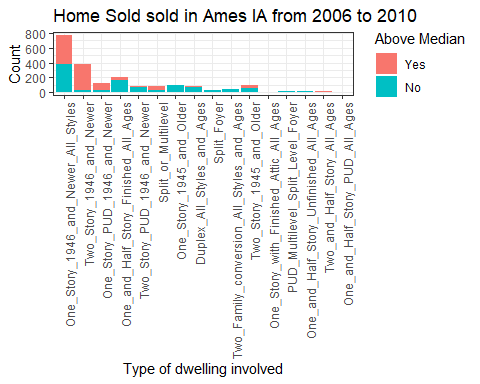
Larger basement houses tend to have above median price.

larger first floor houses tend to have above median price.

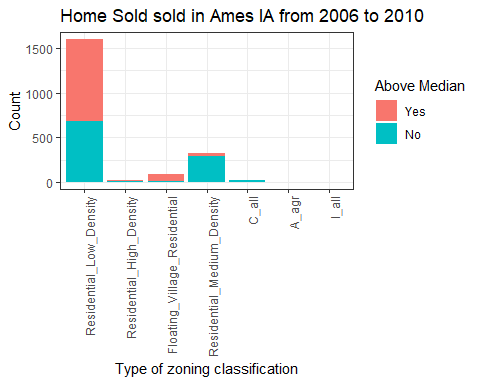
Bigger(gereter square feet) houses tend to have above median price.

Bigger garage houses tend to have above median price.

ggplot(ames, aes(x=MS\_SubClass,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Type of dwelling involved",  
 y = "Count")



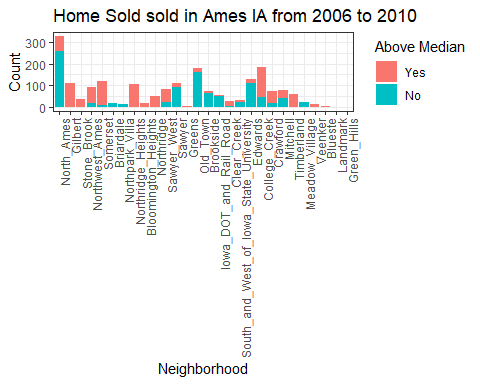
ggplot(ames, aes(x=MS\_Zoning,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Type of zoning classification",  
 y = "Count")



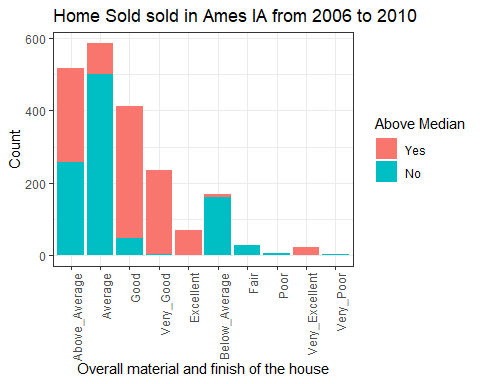
As shown in the bar diagram, two story 1946 and newer houses tend to have above median price.

Also, residental medium density zoning tend to have above median price

ggplot(ames, aes(x=Neighborhood,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Neighborhood",  
 y = "Count")



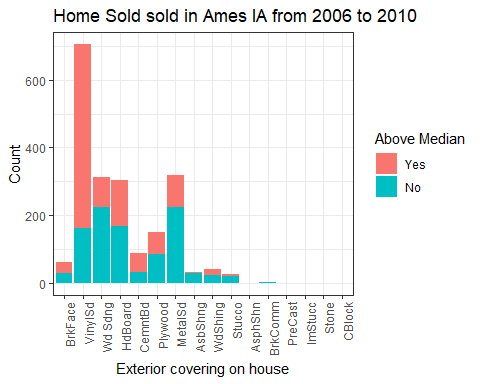
ggplot(ames, aes(x=Overall\_Qual,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Overall material and finish of the house",  
 y = "Count")



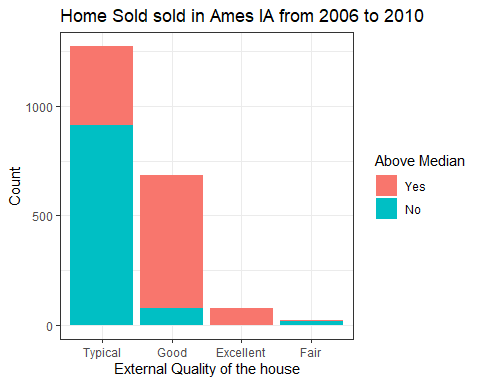
As seen in the Bar diagram (Above), certain neighborhoods have housing prices above median, Example: Northridge Heights, Gilbert, Summerset, Timberland etc.

Overall quality of the house (Left) is very important. Good, very good, excellent and very excellent condition houses tend to have above median prices.

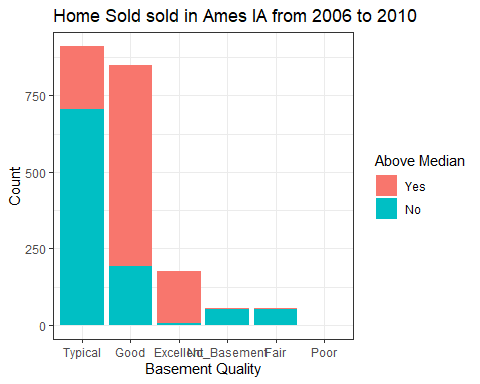
ggplot(ames, aes(x=Exterior\_1st,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Exterior covering on house",  
 y = "Count")

 Exterior covering Vinyl siding houses tend to have above median price.

ggplot(ames, aes(x=Exter\_Qual,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "External Quality of the house",  
 y = "Count")



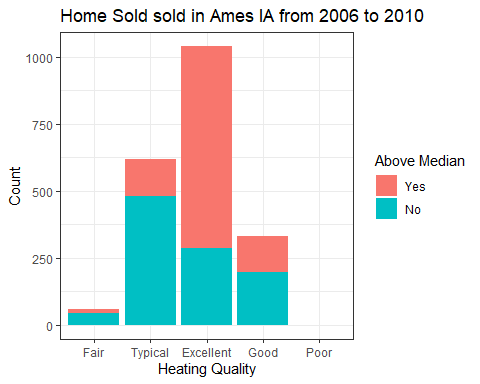
ggplot(ames, aes(x=Bsmt\_Qual,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Basement Quality",  
 y = "Count")



As shown in the diagram, good or excellent external quality of the houses tend to have above median price.

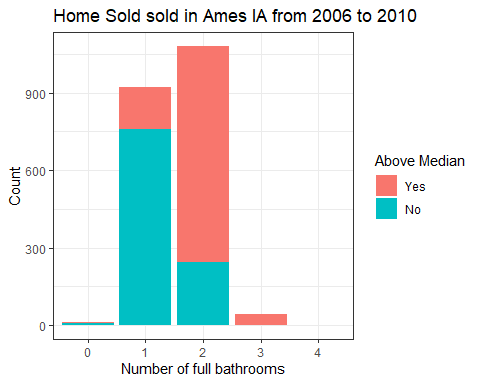
As shown in the diagram, good or excellent basement quality of the houses tend to have above median price.

ggplot(ames, aes(x=Heating\_QC,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Heating Quality",  
 y = "Count")

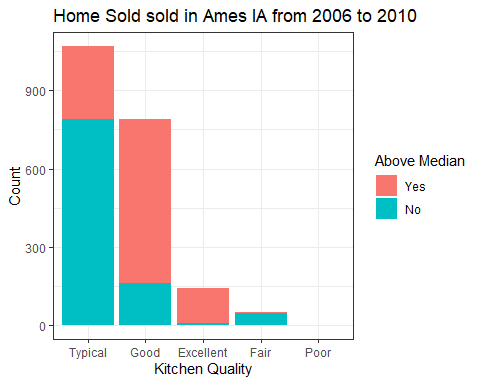


As shown in the diagram above, excellent heating quality of the houses tend to have above median price.

ggplot(ames, aes(x=Full\_Bath,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median")+  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Number of full bathrooms",  
 y = "Count")



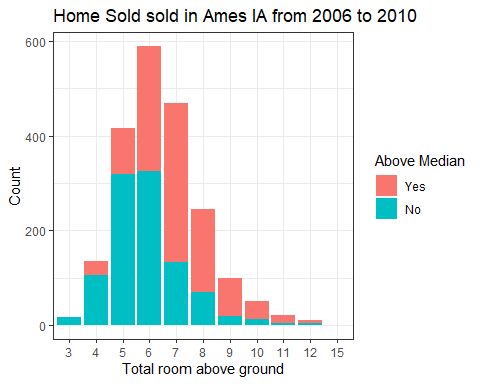
ggplot(ames, aes(x=Kitchen\_Qual,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median")+  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Kitchen Quality",  
 y = "Count")



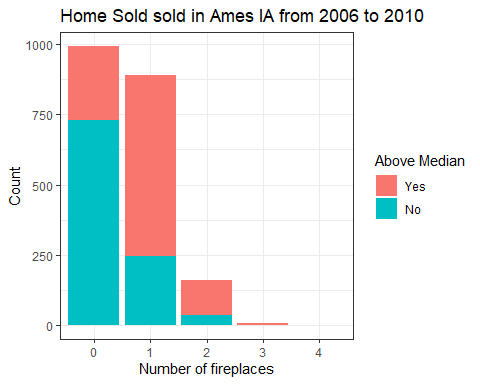
As shown in the bar chart, houses with number of full bathrooms two or more tend to have above median price.

As shown in the diagram above, houses with good or excellent kitchen quality tend to have above median price.

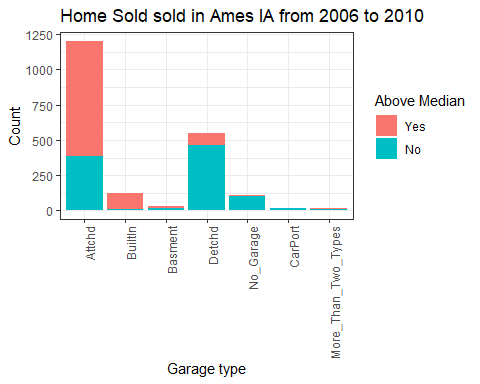
ggplot(ames, aes(x=TotRms\_AbvGrd,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median")+  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Total room above ground",  
 y = "Count")



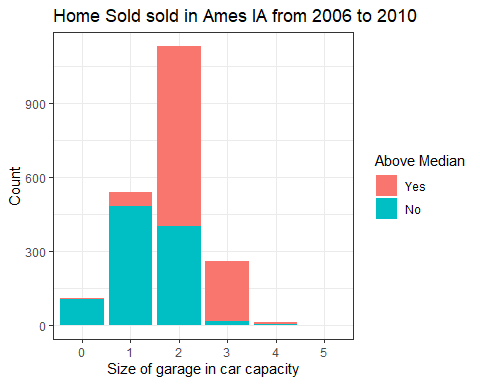
ggplot(ames, aes(x=Fireplaces,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Number of fireplaces",  
 y = "Count")



ggplot(ames, aes(x=Garage\_Type,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 theme(axis.text.x=element\_text(angle = 90, hjust=1)) +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Garage type",  
 y = "Count")



ggplot(ames, aes(x=Garage\_Cars,fill=Above\_Median)) + geom\_bar() + theme\_bw() +  
 scale\_fill\_discrete(name="Above Median") +  
 labs(title = "Home Sold sold in Ames IA from 2006 to 2010",   
 x = "Size of garage in car capacity",  
 y = "Count")



Number of rooms which reflects size of the house. As the number of total rooms above ground increases there is higher chances of housing prices to be above median.

One or more fireplaces houses tend to have to have above median price than houses with none fireplace.

Houses with attached and built in garage tend to have to have above median price than houses with none fireplace.

As seen in the Bar diagram (Above), size of car garage matters. The bigger garage (more car capacity), the greater chance of housing prices to be above median.

Home buying process is stressful specially for first time home buyers. So, they have to make sure they are paying reasonable amount compared to others. For example, you probably do not want to pay above median home value in old town neighborhood. So, this dataset Ames is very useful for Ames, Iowa home buyers.

When analyzing thoroughly variables like original construction year, remodeling year, size of the house, condition of the house, the neighborhood, quality of the material/appliances used, zoning, HVAC quality, Exterior covering and its material quality, number of bathrooms, Garage capacity and type, fireplaces, kitchen quality seem to make more influence on the housing prices.   It is recommended to look into all the variables that influence more in the housing prices. In addition to mentioned variables, one should check crime rate, school ratings, traffic, amenities etc.