library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──  
## ✔ broom 1.0.1 ✔ rsample 1.1.0  
## ✔ dials 1.0.0 ✔ tune 1.0.0  
## ✔ infer 1.0.3 ✔ workflows 1.0.0  
## ✔ modeldata 1.0.1 ✔ workflowsets 1.0.0  
## ✔ parsnip 1.0.1 ✔ yardstick 1.0.0  
## ✔ recipes 1.0.1   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Use suppressPackageStartupMessages() to eliminate package startup messages

library(esquisse)  
library(GGally)

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

library(ggcorrplot)  
library(gridExtra) #I may need to put plots in a grid for easier viewing

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

library(leaps)  
library(skimr)

#Read in the dataset

ames\_homesales <- read\_csv("ames\_homesales.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.

#skim(ames\_homesales) #Good news, no missing data  
#summary(ames\_homesales)

ames\_homesales[sapply(ames\_homesales, is.character)] <- lapply(ames\_homesales[sapply(ames\_homesales, is.character)],   
 as.factor)  
  
#Response Variable, Above\_Median, is a character vector. Also, let's convert all the variables that are characters to factors

str(ames\_homesales)

## spec\_tbl\_df [2,053 × 81] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ MS\_SubClass : Factor w/ 16 levels "Duplex\_All\_Styles\_and\_Ages",..: 6 6 6 6 15 15 7 7 6 6 ...  
## $ MS\_Zoning : Factor w/ 7 levels "A\_agr","C\_all",..: 6 5 6 6 6 6 6 6 6 6 ...  
## $ Lot\_Frontage : num [1:2053] 141 80 81 93 74 78 43 39 0 85 ...  
## $ Lot\_Area : num [1:2053] 31770 11622 14267 11160 13830 ...  
## $ Street : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Alley : Factor w/ 3 levels "Gravel","No\_Alley\_Access",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Lot\_Shape : Factor w/ 4 levels "Irregular","Moderately\_Irregular",..: 4 3 4 3 4 4 4 4 4 3 ...  
## $ Land\_Contour : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 2 4 4 4 ...  
## $ Utilities : Factor w/ 2 levels "AllPub","NoSewr": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Lot\_Config : Factor w/ 5 levels "Corner","CulDSac",..: 1 5 1 1 5 5 5 5 5 5 ...  
## $ Land\_Slope : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Neighborhood : Factor w/ 28 levels "Bloomington\_Heights",..: 16 16 16 16 9 9 26 26 9 9 ...  
## $ Condition\_1 : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 3 3 3 ...  
## $ Condition\_2 : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ Bldg\_Type : Factor w/ 5 levels "Duplex","OneFam",..: 2 2 2 2 2 2 4 4 2 2 ...  
## $ House\_Style : Factor w/ 8 levels "One\_and\_Half\_Fin",..: 3 3 3 3 8 8 3 3 3 3 ...  
## $ Overall\_Qual : Factor w/ 10 levels "Above\_Average",..: 1 2 1 6 2 1 9 9 1 6 ...  
## $ Overall\_Cond : Factor w/ 9 levels "Above\_Average",..: 2 1 1 2 2 1 2 2 6 2 ...  
## $ Year\_Built : num [1:2053] 1960 1961 1958 1968 1997 ...  
## $ Year\_Remod\_Add : num [1:2053] 1960 1961 1958 1968 1998 ...  
## $ Roof\_Style : Factor w/ 6 levels "Flat","Gable",..: 4 2 4 4 2 2 2 2 2 2 ...  
## $ Roof\_Matl : Factor w/ 6 levels "CompShg","Metal",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Exterior\_1st : Factor w/ 16 levels "AsbShng","AsphShn",..: 4 14 15 4 14 14 7 6 7 7 ...  
## $ Exterior\_2nd : Factor w/ 17 levels "AsbShng","AsphShn",..: 11 15 16 4 15 15 7 6 7 7 ...  
## $ Mas\_Vnr\_Type : Factor w/ 5 levels "BrkCmn","BrkFace",..: 5 4 2 4 4 2 4 4 4 4 ...  
## $ Mas\_Vnr\_Area : num [1:2053] 112 0 108 0 0 20 0 0 0 0 ...  
## $ Exter\_Qual : Factor w/ 4 levels "Excellent","Fair",..: 4 4 4 3 4 4 3 3 4 4 ...  
## $ Exter\_Cond : Factor w/ 5 levels "Excellent","Fair",..: 5 5 5 5 5 5 5 5 3 5 ...  
## $ Foundation : Factor w/ 6 levels "BrkTil","CBlock",..: 2 2 2 2 3 3 3 3 3 3 ...  
## $ Bsmt\_Qual : Factor w/ 6 levels "Excellent","Fair",..: 6 6 6 6 3 6 3 3 3 3 ...  
## $ Bsmt\_Cond : Factor w/ 6 levels "Excellent","Fair",..: 3 6 6 6 6 6 6 6 6 6 ...  
## $ Bsmt\_Exposure : Factor w/ 5 levels "Av","Gd","Mn",..: 2 4 4 4 4 4 4 4 4 2 ...  
## $ BsmtFin\_Type\_1 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 2 6 1 1 3 3 1 3 1 3 ...  
## $ BsmtFin\_SF\_1 : num [1:2053] 2 6 1 1 3 3 1 3 1 3 ...  
## $ BsmtFin\_Type\_2 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 7 4 7 7 7 7 7 7 7 7 ...  
## $ BsmtFin\_SF\_2 : num [1:2053] 0 144 0 0 0 0 0 0 0 0 ...  
## $ Bsmt\_Unf\_SF : num [1:2053] 441 270 406 1045 137 ...  
## $ Total\_Bsmt\_SF : num [1:2053] 1080 882 1329 2110 928 ...  
## $ Heating : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Heating\_QC : Factor w/ 5 levels "Excellent","Fair",..: 2 5 5 1 3 1 1 1 1 3 ...  
## $ Central\_Air : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Electrical : Factor w/ 5 levels "FuseA","FuseF",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ First\_Flr\_SF : num [1:2053] 1656 896 1329 2110 928 ...  
## $ Second\_Flr\_SF : num [1:2053] 0 0 0 0 701 678 0 0 0 0 ...  
## $ Low\_Qual\_Fin\_SF : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Gr\_Liv\_Area : num [1:2053] 1656 896 1329 2110 1629 ...  
## $ Bsmt\_Full\_Bath : num [1:2053] 1 0 0 1 0 0 0 1 1 1 ...  
## $ Bsmt\_Half\_Bath : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Full\_Bath : num [1:2053] 1 1 1 2 2 2 2 2 2 1 ...  
## $ Half\_Bath : num [1:2053] 0 0 1 1 1 1 0 0 0 1 ...  
## $ Bedroom\_AbvGr : num [1:2053] 3 2 3 3 3 3 2 2 3 2 ...  
## $ Kitchen\_AbvGr : num [1:2053] 1 1 1 1 1 1 1 1 1 1 ...  
## $ Kitchen\_Qual : Factor w/ 5 levels "Excellent","Fair",..: 5 5 3 1 5 3 3 3 5 3 ...  
## $ TotRms\_AbvGrd : num [1:2053] 7 5 6 8 6 7 5 5 6 5 ...  
## $ Functional : Factor w/ 8 levels "Maj1","Maj2",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ Fireplaces : num [1:2053] 2 0 0 2 1 1 0 1 0 1 ...  
## $ Fireplace\_Qu : Factor w/ 6 levels "Excellent","Fair",..: 3 4 4 6 6 3 4 6 4 5 ...  
## $ Garage\_Type : Factor w/ 7 levels "Attchd","Basment",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Garage\_Finish : Factor w/ 4 levels "Fin","No\_Garage",..: 1 4 4 1 1 1 3 3 1 4 ...  
## $ Garage\_Cars : num [1:2053] 2 1 1 2 2 2 2 2 2 2 ...  
## $ Garage\_Area : num [1:2053] 528 730 312 522 482 470 506 608 420 506 ...  
## $ Garage\_Qual : Factor w/ 6 levels "Excellent","Fair",..: 6 6 6 6 6 6 6 6 6 6 ...  
## $ Garage\_Cond : Factor w/ 6 levels "Excellent","Fair",..: 6 6 6 6 6 6 6 6 6 6 ...  
## $ Paved\_Drive : Factor w/ 3 levels "Dirt\_Gravel",..: 2 3 3 3 3 3 3 3 3 3 ...  
## $ Wood\_Deck\_SF : num [1:2053] 210 140 393 0 212 360 0 237 483 192 ...  
## $ Open\_Porch\_SF : num [1:2053] 62 0 36 0 34 36 82 152 21 0 ...  
## $ Enclosed\_Porch : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Three\_season\_porch: num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Screen\_Porch : num [1:2053] 0 120 0 0 0 0 144 0 0 0 ...  
## $ Pool\_Area : num [1:2053] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Pool\_QC : Factor w/ 5 levels "Excellent","Fair",..: 4 4 4 4 4 4 4 4 4 4 ...  
## $ Fence : Factor w/ 5 levels "Good\_Privacy",..: 5 3 5 5 3 5 5 5 1 5 ...  
## $ Misc\_Feature : Factor w/ 5 levels "Elev","Gar2",..: 3 3 2 3 3 3 3 3 5 3 ...  
## $ Misc\_Val : num [1:2053] 0 0 12500 0 0 0 0 0 500 0 ...  
## $ Mo\_Sold : num [1:2053] 5 6 6 4 3 6 1 3 3 2 ...  
## $ Year\_Sold : num [1:2053] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...  
## $ Sale\_Type : Factor w/ 10 levels "COD","Con","ConLD",..: 10 10 10 10 10 10 10 10 10 10 ...  
## $ Sale\_Condition : Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 5 5 5 5 5 5 5 ...  
## $ Longitude : num [1:2053] -93.6 -93.6 -93.6 -93.6 -93.6 ...  
## $ Latitude : num [1:2053] 42.1 42.1 42.1 42.1 42.1 ...  
## $ Above\_Median : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. MS\_SubClass = col\_character(),  
## .. MS\_Zoning = col\_character(),  
## .. Lot\_Frontage = col\_double(),  
## .. Lot\_Area = col\_double(),  
## .. Street = col\_character(),  
## .. Alley = col\_character(),  
## .. Lot\_Shape = col\_character(),  
## .. Land\_Contour = col\_character(),  
## .. Utilities = col\_character(),  
## .. Lot\_Config = col\_character(),  
## .. Land\_Slope = col\_character(),  
## .. Neighborhood = col\_character(),  
## .. Condition\_1 = col\_character(),  
## .. Condition\_2 = col\_character(),  
## .. Bldg\_Type = col\_character(),  
## .. House\_Style = col\_character(),  
## .. Overall\_Qual = col\_character(),  
## .. Overall\_Cond = col\_character(),  
## .. Year\_Built = col\_double(),  
## .. Year\_Remod\_Add = col\_double(),  
## .. Roof\_Style = col\_character(),  
## .. Roof\_Matl = col\_character(),  
## .. Exterior\_1st = col\_character(),  
## .. Exterior\_2nd = col\_character(),  
## .. Mas\_Vnr\_Type = col\_character(),  
## .. Mas\_Vnr\_Area = col\_double(),  
## .. Exter\_Qual = col\_character(),  
## .. Exter\_Cond = col\_character(),  
## .. Foundation = col\_character(),  
## .. Bsmt\_Qual = col\_character(),  
## .. Bsmt\_Cond = col\_character(),  
## .. Bsmt\_Exposure = col\_character(),  
## .. BsmtFin\_Type\_1 = col\_character(),  
## .. BsmtFin\_SF\_1 = col\_double(),  
## .. BsmtFin\_Type\_2 = col\_character(),  
## .. BsmtFin\_SF\_2 = col\_double(),  
## .. Bsmt\_Unf\_SF = col\_double(),  
## .. Total\_Bsmt\_SF = col\_double(),  
## .. Heating = col\_character(),  
## .. Heating\_QC = col\_character(),  
## .. Central\_Air = col\_character(),  
## .. Electrical = col\_character(),  
## .. 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(),  
## .. Kitchen\_Qual = col\_character(),  
## .. TotRms\_AbvGrd = col\_double(),  
## .. Functional = col\_character(),  
## .. Fireplaces = col\_double(),  
## .. Fireplace\_Qu = col\_character(),  
## .. Garage\_Type = col\_character(),  
## .. Garage\_Finish = col\_character(),  
## .. Garage\_Cars = col\_double(),  
## .. Garage\_Area = col\_double(),  
## .. Garage\_Qual = col\_character(),  
## .. Garage\_Cond = col\_character(),  
## .. Paved\_Drive = col\_character(),  
## .. Wood\_Deck\_SF = col\_double(),  
## .. Open\_Porch\_SF = col\_double(),  
## .. Enclosed\_Porch = col\_double(),  
## .. Three\_season\_porch = col\_double(),  
## .. Screen\_Porch = col\_double(),  
## .. Pool\_Area = col\_double(),  
## .. Pool\_QC = col\_character(),  
## .. Fence = col\_character(),  
## .. Misc\_Feature = col\_character(),  
## .. Misc\_Val = col\_double(),  
## .. Mo\_Sold = col\_double(),  
## .. Year\_Sold = col\_double(),  
## .. Sale\_Type = col\_character(),  
## .. Sale\_Condition = col\_character(),  
## .. Longitude = col\_double(),  
## .. Latitude = col\_double(),  
## .. Above\_Median = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

ames <- ames\_homesales %>%  
select(c(1:78,81))  
#Removing columns 79 & 80 (Longitude and Latitude) to reduce the complexity since these are variables that don't add to our predictions.

table(ames$Above\_Median) #1010 properties not above median, 1043 above median

##   
## No Yes   
## 1010 1043

#this is the response variable and the dataset appears balanced with only slightly more than half of the properties selling above median

t1 = table(ames$Above\_Median, ames$Year\_Sold) #create a table object  
prop.table(t1, margin = 2 ) #crosstab with proportions

##   
## 2006 2007 2008 2009 2010  
## No 0.4909502 0.4669339 0.4966292 0.5043860 0.5165877  
## Yes 0.5090498 0.5330661 0.5033708 0.4956140 0.4834123

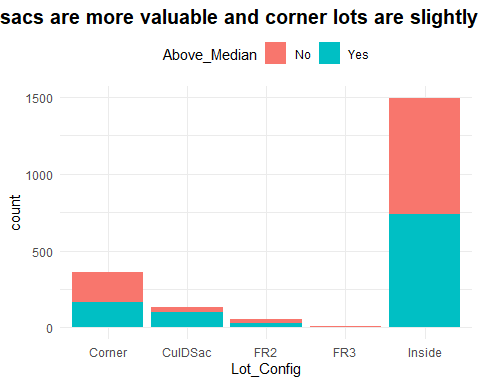
#Proportion of homes selling above (or below) median (response variable) by year sold

t2 = table(ames$Above\_Median, ames$Lot\_Config)  
prop.table(t2, margin = 2)

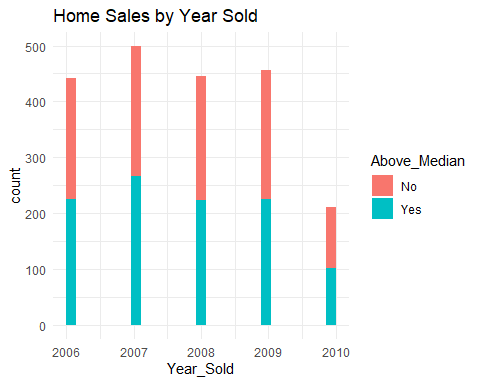
##   
## Corner CulDSac FR2 FR3 Inside  
## No 0.5320334 0.2518519 0.4821429 0.3750000 0.5050167  
## Yes 0.4679666 0.7481481 0.5178571 0.6250000 0.4949833

#Proportion of properties selling above median by lot config. CulDeSacs are selling above median, as are FR3.

ggplot(ames) +  
 aes(x = Lot\_Config, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = " Cul-de-sacs are more valuable and corner lots are slightly less valuable") +  
 theme\_minimal() +  
 theme(legend.position = "top", plot.title = element\_text(size = 15L, face = "bold",   
 hjust = 0.5), plot.subtitle = element\_text(size = 12L, hjust = 0.5))

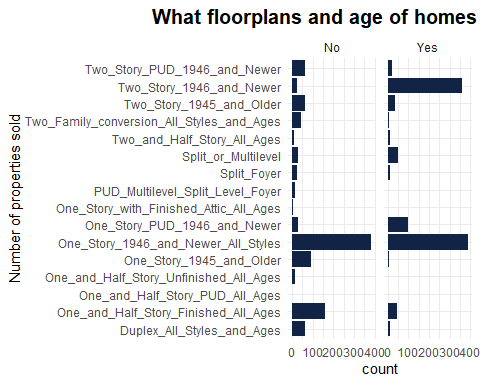


#How many properties sold each year?  
ggplot(ames) +  
 aes(x = Year\_Sold, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Home Sales by Year Sold") +  
 theme\_minimal()

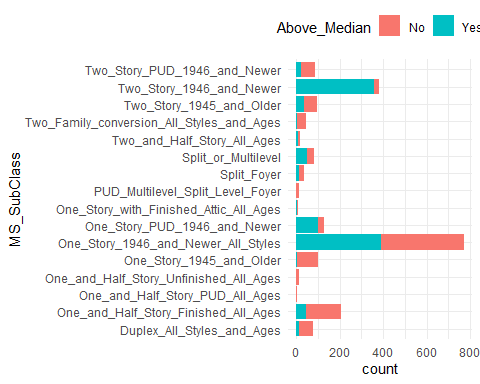


#We learn that sales dropped in 2010 but as we already knew from the table, proportion of above and below median was uniform across the years.

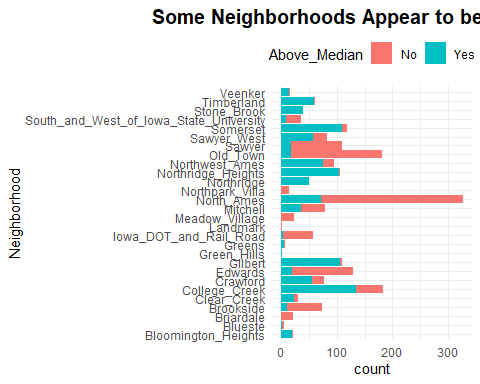
#esquisser  
  
  
ggplot(ames) +  
 aes(x = MS\_SubClass) +  
 geom\_bar(fill = "#112446") +  
 labs(x = "Number of properties sold",   
 title = "What floorplans and age of homes were selling?") +  
 coord\_flip() +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 15L,   
 face = "bold", hjust = 0.5), plot.subtitle = element\_text(size = 12L, hjust = 0.5)) +  
 facet\_wrap(vars(Above\_Median))



ggplot(ames) +  
 aes(x = MS\_SubClass, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 coord\_flip() +  
 theme\_minimal() +  
 theme(legend.position = "top", plot.title = element\_text(size = 15L,   
 face = "bold", hjust = 0.5), plot.subtitle = element\_text(size = 12L, hjust = 0.5))



ggplot(ames) +  
 aes(x = Neighborhood, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Some Neighborhoods Appear to be Hot Sellers") +  
 coord\_flip() +  
 theme\_minimal() +  
 theme(legend.position = "top",   
 plot.title = element\_text(size = 15L, face = "bold", hjust = 0.5), plot.subtitle = element\_text(size = 12L,   
 hjust = 0.5))



table(ames$Street) #Only 7 properties in this category are in "gravel" category, all others are "street"

##   
## Grvl Pave   
## 7 2046

table(ames$Utilities) #Only 1 property is "no sewr" all other are "public"

##   
## AllPub NoSewr   
## 2052 1

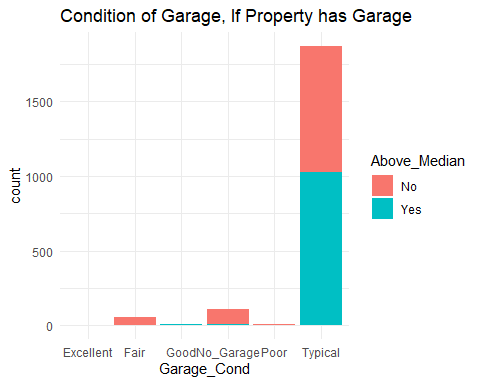
table(ames$Lot\_Shape)

##   
## Irregular Moderately\_Irregular Regular   
## 11 53 1275   
## Slightly\_Irregular   
## 714

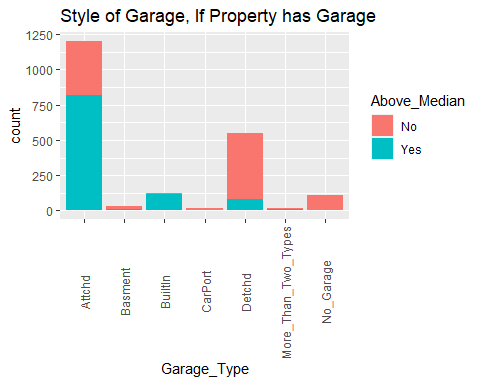
table(ames$Half\_Bath) #1300 have no half bath, 736 have 1 half bath and 17 have 2 half baths

##   
## 0 1 2   
## 1300 736 17

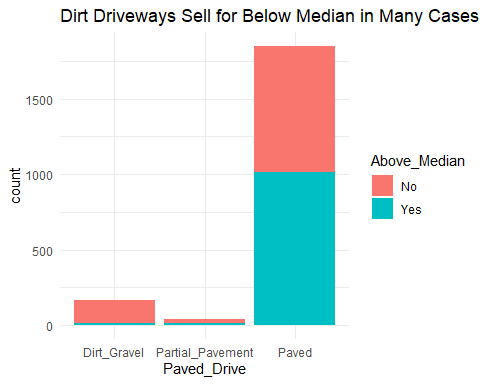
ggplot(ames) +  
 aes(x = Garage\_Cond, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Condition of Garage, If Property has Garage") +  
 theme\_minimal()



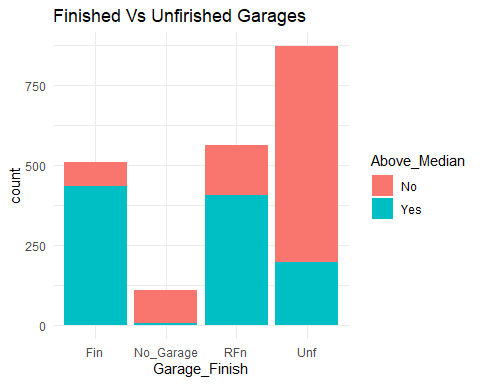
ggplot(ames) +  
 aes(x = Garage\_Type, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Style of Garage, If Property has Garage") +  
 theme(axis.text.x = element\_text(angle = 90,vjust = .5))



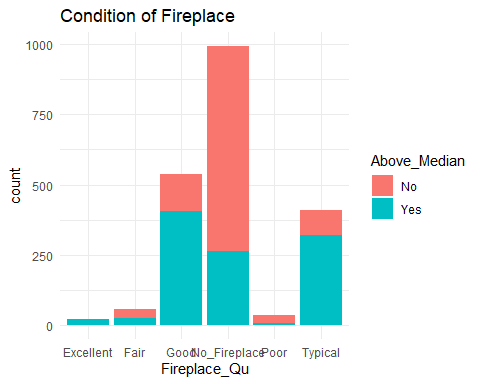
ggplot(ames) +  
 aes(x = Paved\_Drive, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Dirt Driveways Sell for Below Median in Many Cases") +  
 theme\_minimal()



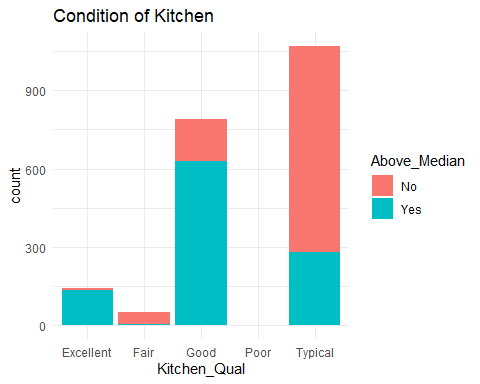
ggplot(ames) +  
 aes(x = Garage\_Finish, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Finished Vs Unfirished Garages") +  
 theme\_minimal()



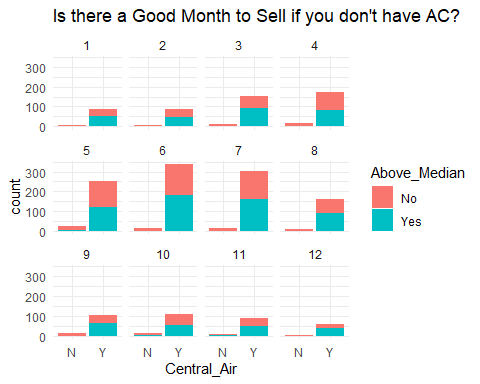
ggplot(ames) +  
 aes(x = Fireplace\_Qu, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Condition of Fireplace") +  
 theme\_minimal()



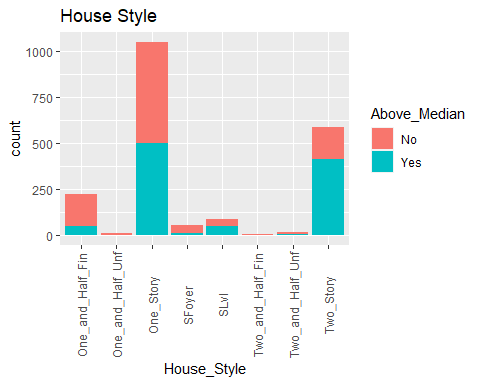
ggplot(ames) +  
 aes(x = Kitchen\_Qual, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Condition of Kitchen") +  
 theme\_minimal()



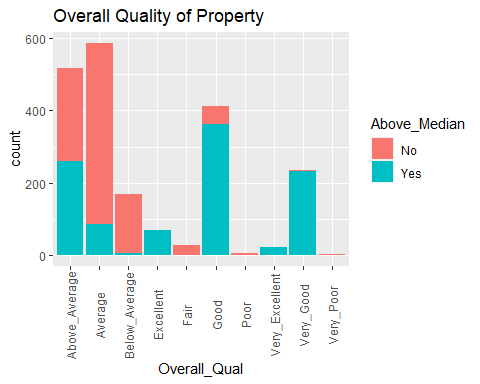
ggplot(ames) +  
 aes(x = Central\_Air, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Is there a Good Month to Sell if you don't have AC?") +  
 theme\_minimal() +  
 facet\_wrap(vars(Mo\_Sold))



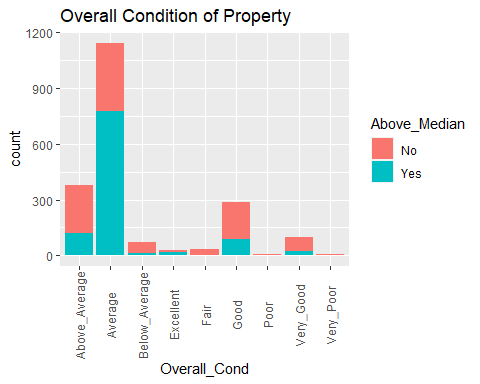
ggplot(ames) +  
 aes(x = House\_Style, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "House Style") +  
 theme(axis.text.x = element\_text(angle = 90,vjust = .5))



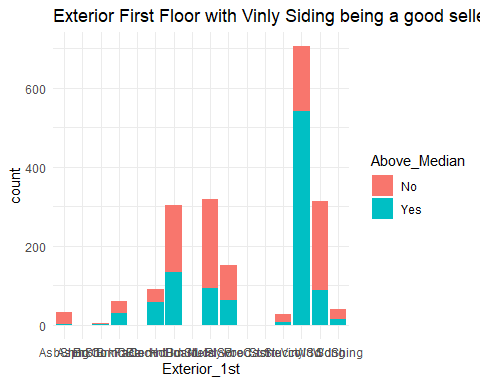
ggplot(ames) +  
 aes(x = Overall\_Qual, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Overall Quality of Property") +  
 theme(axis.text.x = element\_text(angle = 90,vjust = .5))



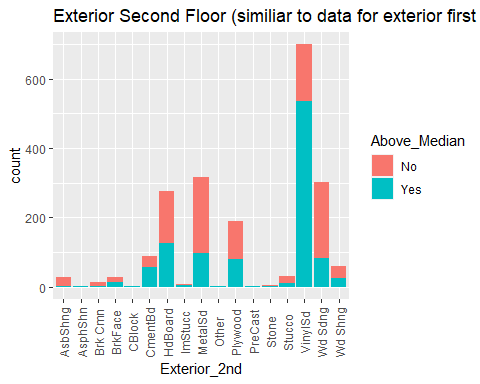
ggplot(ames) +  
 aes(x = Overall\_Cond, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Overall Condition of Property") +  
 theme(axis.text.x = element\_text(angle = 90,vjust = .5))



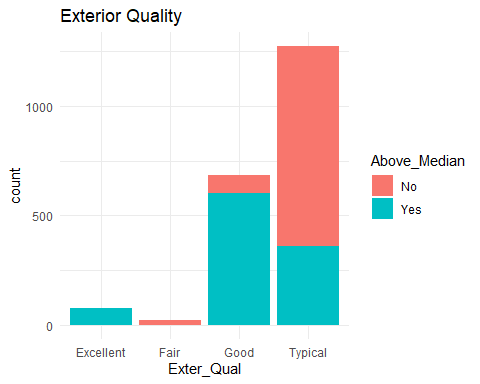
ggplot(ames) +  
 aes(x = Exterior\_1st, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Exterior First Floor with Vinly Siding being a good seller") +  
 theme\_minimal()



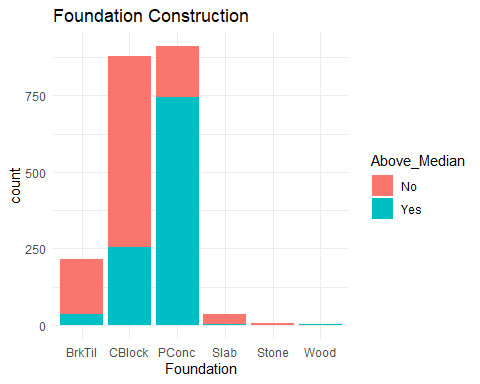
ggplot(ames) +  
 aes(x = Exterior\_2nd, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Exterior Second Floor (similiar to data for exterior first floor)") +  
 theme(axis.text.x = element\_text(angle = 90,vjust = .5))



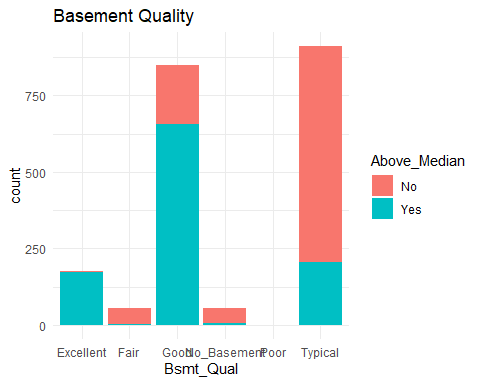
ggplot(ames) +  
 aes(x = Exter\_Qual, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Exterior Quality") +  
 theme\_minimal()



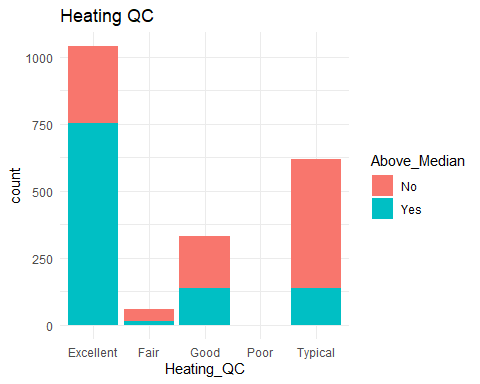
ggplot(ames) +  
 aes(x = Foundation, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Foundation Construction") +  
 theme\_minimal()



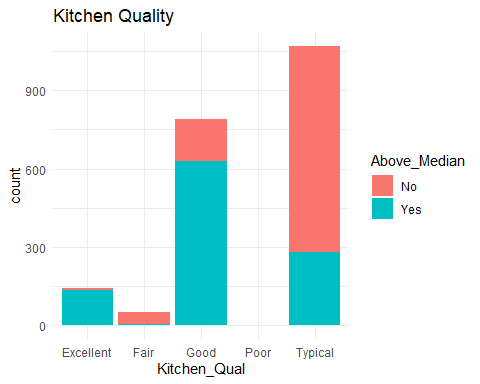
ggplot(ames) +  
 aes(x = Bsmt\_Qual, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Basement Quality") +  
 theme\_minimal()



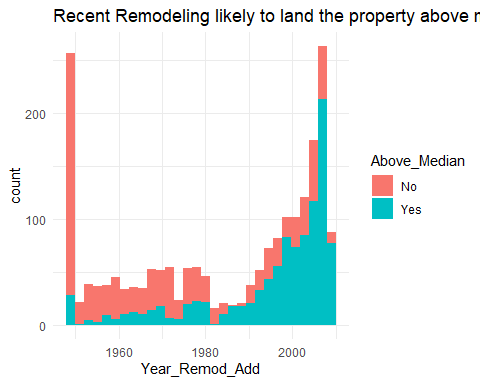
ggplot(ames) +  
 aes(x = Heating\_QC, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Heating QC") +  
 theme\_minimal()



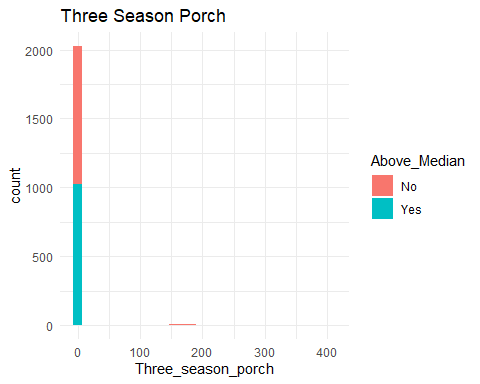
ggplot(ames) +  
 aes(x = Kitchen\_Qual, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Kitchen Quality") +  
 theme\_minimal()



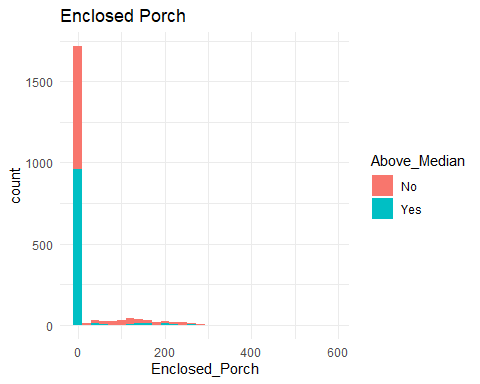
ggplot(ames) +  
 aes(x = Year\_Remod\_Add, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Recent Remodeling likely to land the property above median") +  
 theme\_minimal()



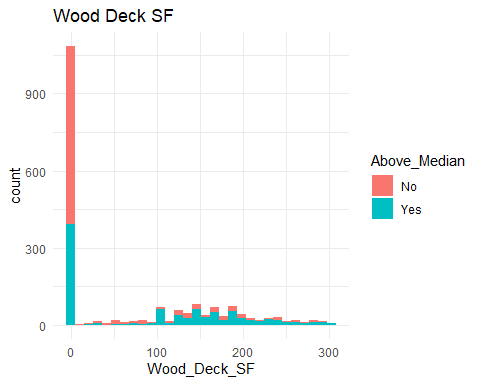
ggplot(ames) +  
 aes(x = Three\_season\_porch, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Three Season Porch") +  
 theme\_minimal()



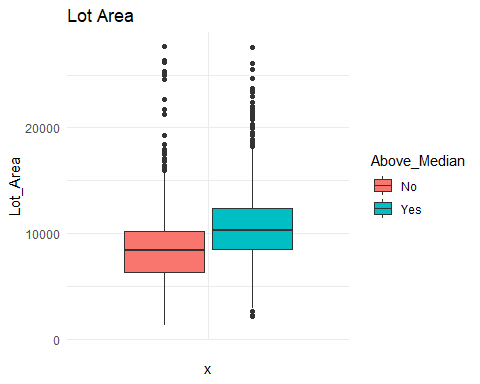
ggplot(ames) +  
 aes(x = Enclosed\_Porch, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Enclosed Porch") +  
 theme\_minimal()



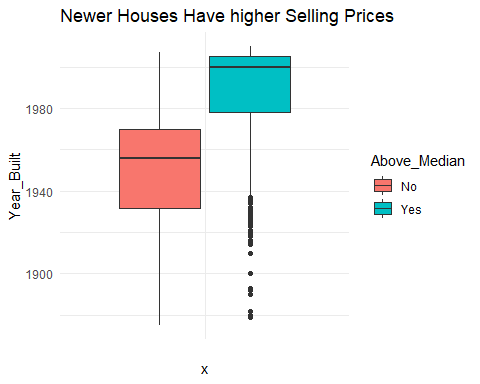
ames %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <= 302L) %>%  
 ggplot() +  
 aes(x = Wood\_Deck\_SF, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Wood Deck SF") +  
 theme\_minimal()



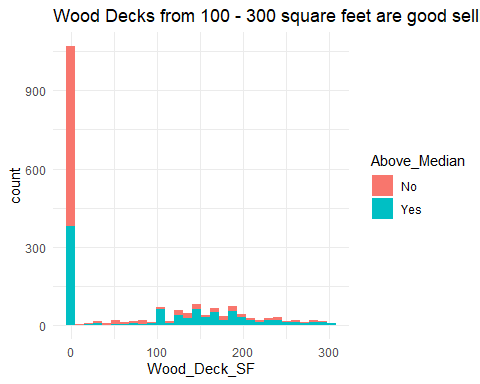
ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <=   
 302L) %>%  
 ggplot() +  
 aes(x = "", y = Lot\_Area, fill = Above\_Median) +  
 geom\_boxplot() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Lot Area") +  
 theme\_minimal()



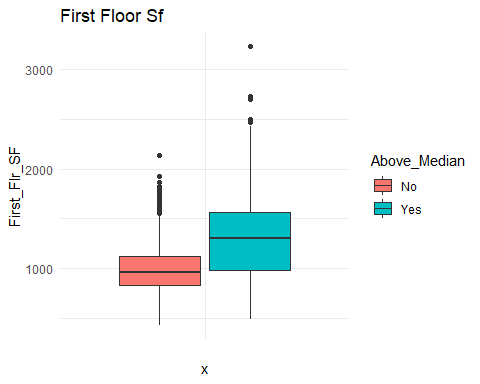
ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <=   
 302L) %>%  
 ggplot() +  
 aes(x = "", y = Year\_Built, fill = Above\_Median) +  
 geom\_boxplot() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Newer Houses Have higher Selling Prices") +  
 theme\_minimal()



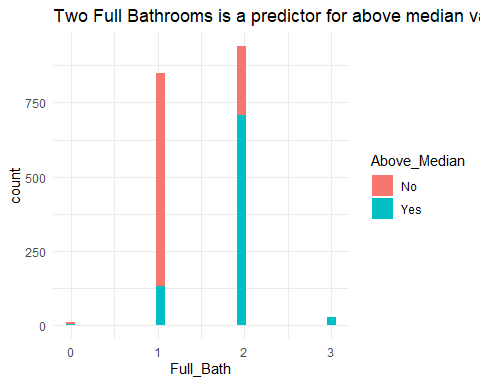
ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <=   
 302L) %>%  
 ggplot() +  
 aes(x = Wood\_Deck\_SF, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Wood Decks from 100 - 300 square feet are good selling points") +  
 theme\_minimal()



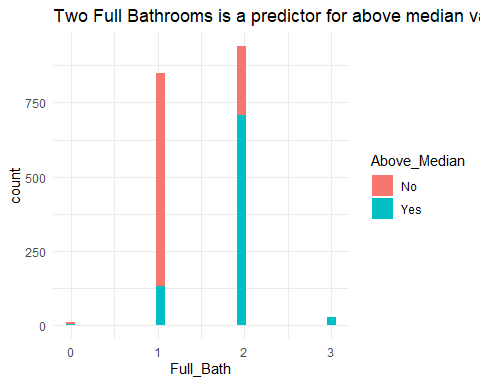
ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <=   
 302L) %>%  
 ggplot() +  
 aes(x = "", y = First\_Flr\_SF, fill = Above\_Median) +  
 geom\_boxplot() +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "First Floor Sf") +  
 theme\_minimal()



ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Mas\_Vnr\_Area >= 0L & Mas\_Vnr\_Area <=   
 502L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <= 302L) %>%  
 ggplot() +  
 aes(x = Full\_Bath, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Two Full Bathrooms is a predictor for above median value") +  
 theme\_minimal()



ames %>%  
 filter(Lot\_Area >= 1300L & Lot\_Area <= 28280L) %>%  
 filter(Mas\_Vnr\_Area >= 0L & Mas\_Vnr\_Area <=   
 502L) %>%  
 filter(Wood\_Deck\_SF >= 0L & Wood\_Deck\_SF <= 302L) %>%  
 ggplot() +  
 aes(x = Full\_Bath, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(title = "Two Full Bathrooms is a predictor for above median value") +  
 theme\_minimal()

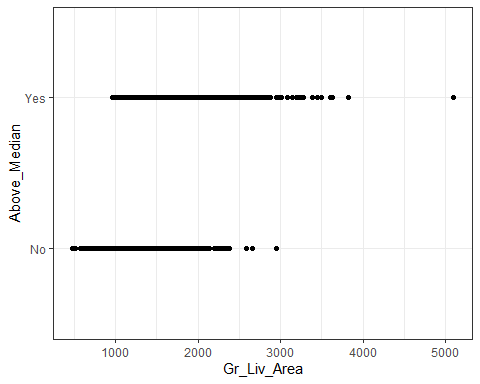


t9 = table(ames$Above\_Median, ames$Full\_Bath)  
prop.table(t9, margin = 2)

##   
## 0 1 2 3 4  
## No 0.70000000 0.82500000 0.22500000 0.02439024 0.00000000  
## Yes 0.30000000 0.17500000 0.77500000 0.97560976 1.00000000

#ggcorr(ames) #Correlation is not calculated for non-quantitative variables #ggcorr ignores non-numeric variables #Heat map indicates: Garage Cars, Bedroom AbvGr, Gr\_Liv\_Area, Total Bsmt SF, Year Built, Bsmt fin Sq Ft 1 and Bsmt Unfin are all correlated with each other however, our response variable isn’t even in this equation because it is a categorical variable. So, basically no predictive response ntoed.

ggplot(ames, aes(x=Gr\_Liv\_Area, y=Above\_Median)) + geom\_point() + theme\_bw()



#Some outliers are homes with large square footage, above 3000 sq feet

Ames\_MS\_SubClass <- ames\_homesales %>%  
 group\_by(MS\_SubClass) %>%  
 summarize(freq = n()) %>%  
 arrange(desc(freq))  
  
Ames\_MS\_SubClass

## # A tibble: 16 × 2  
## MS\_SubClass freq  
## <fct> <int>  
## 1 One\_Story\_1946\_and\_Newer\_All\_Styles 772  
## 2 Two\_Story\_1946\_and\_Newer 383  
## 3 One\_and\_Half\_Story\_Finished\_All\_Ages 204  
## 4 One\_Story\_PUD\_1946\_and\_Newer 129  
## 5 One\_Story\_1945\_and\_Older 98  
## 6 Two\_Story\_1945\_and\_Older 95  
## 7 Two\_Story\_PUD\_1946\_and\_Newer 85  
## 8 Split\_or\_Multilevel 82  
## 9 Duplex\_All\_Styles\_and\_Ages 76  
## 10 Two\_Family\_conversion\_All\_Styles\_and\_Ages 46  
## 11 Split\_Foyer 34  
## 12 Two\_and\_Half\_Story\_All\_Ages 17  
## 13 One\_and\_Half\_Story\_Unfinished\_All\_Ages 13  
## 14 PUD\_Multilevel\_Split\_Level\_Foyer 12  
## 15 One\_Story\_with\_Finished\_Attic\_All\_Ages 6  
## 16 One\_and\_Half\_Story\_PUD\_All\_Ages 1

t3 = table(ames$Above\_Median, ames$MS\_SubClass) #create a table object  
t3

##   
## Duplex\_All\_Styles\_and\_Ages One\_and\_Half\_Story\_Finished\_All\_Ages  
## No 63 159  
## Yes 13 45  
##   
## One\_and\_Half\_Story\_PUD\_All\_Ages One\_and\_Half\_Story\_Unfinished\_All\_Ages  
## No 1 13  
## Yes 0 0  
##   
## One\_Story\_1945\_and\_Older One\_Story\_1946\_and\_Newer\_All\_Styles  
## No 94 383  
## Yes 4 389  
##   
## One\_Story\_PUD\_1946\_and\_Newer One\_Story\_with\_Finished\_Attic\_All\_Ages  
## No 29 5  
## Yes 100 1  
##   
## PUD\_Multilevel\_Split\_Level\_Foyer Split\_Foyer Split\_or\_Multilevel  
## No 12 23 31  
## Yes 0 11 51  
##   
## Two\_and\_Half\_Story\_All\_Ages Two\_Family\_conversion\_All\_Styles\_and\_Ages  
## No 8 41  
## Yes 9 5  
##   
## Two\_Story\_1945\_and\_Older Two\_Story\_1946\_and\_Newer  
## No 62 23  
## Yes 33 360  
##   
## Two\_Story\_PUD\_1946\_and\_Newer  
## No 63  
## Yes 22

#28 Neighborhoods are represented. Some neighborhoods have very few (only 1) home sale while others have many (327)  
Ames\_gp\_neighborhoods <- ames\_homesales %>%  
 group\_by(Neighborhood) %>%  
 summarize(freq = n()) %>%  
 arrange(desc(freq))  
  
Ames\_gp\_neighborhoods

## # A tibble: 28 × 2  
## Neighborhood freq  
## <fct> <int>  
## 1 North\_Ames 327  
## 2 College\_Creek 183  
## 3 Old\_Town 181  
## 4 Edwards 129  
## 5 Somerset 119  
## 6 Gilbert 109  
## 7 Sawyer 109  
## 8 Northridge\_Heights 105  
## 9 Northwest\_Ames 95  
## 10 Sawyer\_West 82  
## # … with 18 more rows

t8 = table(ames$Above\_Median, ames$Neighborhood) #create a table object  
prop.table(t8, margin = 2)

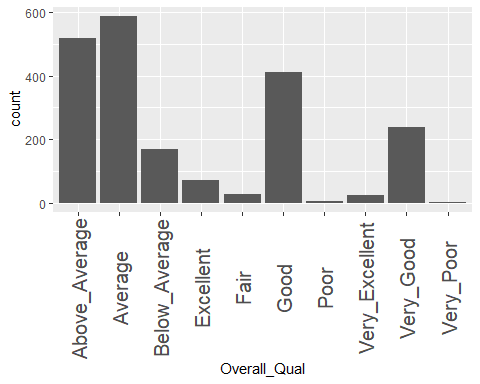
##   
## Bloomington\_Heights Blueste Briardale Brookside Clear\_Creek  
## No 0.09523810 0.80000000 1.00000000 0.86486486 0.22580645  
## Yes 0.90476190 0.20000000 0.00000000 0.13513514 0.77419355  
##   
## College\_Creek Crawford Edwards Gilbert Green\_Hills Greens  
## No 0.26775956 0.27272727 0.84496124 0.02752294 0.00000000 0.14285714  
## Yes 0.73224044 0.72727273 0.15503876 0.97247706 1.00000000 0.85714286  
##   
## Iowa\_DOT\_and\_Rail\_Road Landmark Meadow\_Village Mitchell North\_Ames  
## No 0.92982456 1.00000000 1.00000000 0.54430380 0.77981651  
## Yes 0.07017544 0.00000000 0.00000000 0.45569620 0.22018349  
##   
## Northpark\_Villa Northridge Northridge\_Heights Northwest\_Ames Old\_Town  
## No 1.00000000 0.00000000 0.00952381 0.20000000 0.90055249  
## Yes 0.00000000 1.00000000 0.99047619 0.80000000 0.09944751  
##   
## Sawyer Sawyer\_West Somerset South\_and\_West\_of\_Iowa\_State\_University  
## No 0.84403670 0.30487805 0.07563025 0.74285714  
## Yes 0.15596330 0.69512195 0.92436975 0.25714286  
##   
## Stone\_Brook Timberland Veenker  
## No 0.00000000 0.01666667 0.06250000  
## Yes 1.00000000 0.98333333 0.93750000

#crosstab with proportions  
  
#Proportion of homes selling above (or below) median by neighborhood

Ames\_Overall\_Qual <- ames\_homesales %>%  
 group\_by(Overall\_Qual) %>%  
 summarize(freq = n()) %>%  
 arrange(desc(freq))  
  
Ames\_Overall\_Qual

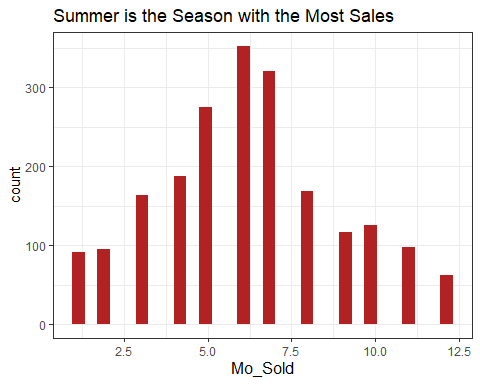
## # A tibble: 10 × 2  
## Overall\_Qual freq  
## <fct> <int>  
## 1 Average 587  
## 2 Above\_Average 518  
## 3 Good 411  
## 4 Very\_Good 237  
## 5 Below\_Average 169  
## 6 Excellent 70  
## 7 Fair 28  
## 8 Very\_Excellent 24  
## 9 Poor 6  
## 10 Very\_Poor 3

ggplot(ames\_homesales, aes(Overall\_Qual))+  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 90, vjust = .5, size = 15))



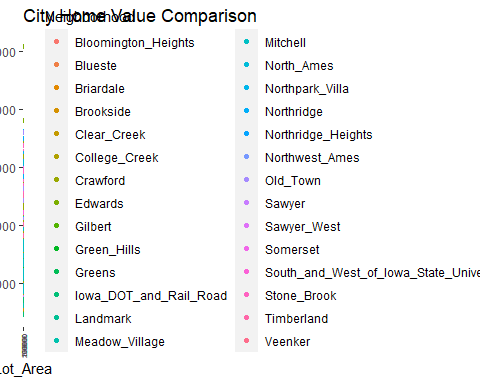
#Looks like some of these conditions are outliers

ggplot(ames) +  
 aes(x = Mo\_Sold) +  
 geom\_histogram(bins = 30L, fill = "#B22222") +  
 labs(title = "Summer is the Season with the Most Sales") +   
 theme\_bw() +  
 theme(axis.title.x = element\_text(size = 12L))

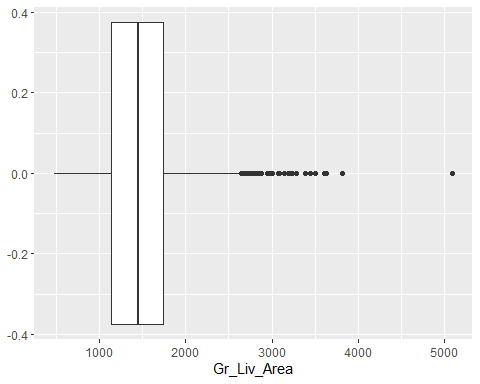


#Let's group the neighborhoods with

ggplot(ames, aes(x = Lot\_Area, y = Gr\_Liv\_Area, color = Neighborhood))+ geom\_point()+   
 labs(title = "City Home Value Comparison", ) +   
 theme(axis.text.x = element\_text(angle = 90, vjust = .5, size = 5))+ scale\_y\_continuous(name = "Living Area Square Footage", labels = scales::comma)



ggplot(ames, aes(Gr\_Liv\_Area))+   
 geom\_boxplot()



table(ames$Gr\_Liv\_Area > 3000) #Sixteen properties are outliers, they are large properties with more than 3000 square feet

##   
## FALSE TRUE   
## 2037 16

#Outliers in living area with 16 of the homes being over 3000 sq feet. Let's remove these 16 observations, which is .77% of the total dataset.  
  
ames <-ames %>%  
 filter(Gr\_Liv\_Area<3000)

t1 <-table(ames$Lot\_Shape)  
  
t2 <- table(ames$MS\_Zoning)  
  
t3 <- table(ames$Bldg\_Type)  
t4 <- table(ames$Roof\_Style)  
t5 <- table(ames$Sale\_Type)  
t6 <- table(ames$Foundation)  
t7 <- table(ames$Land\_Slope)  
  
t1

##   
## Irregular Moderately\_Irregular Regular   
## 11 51 1270   
## Slightly\_Irregular   
## 705

t2

##   
## A\_agr C\_all   
## 2 17   
## Floating\_Village\_Residential I\_all   
## 86 1   
## Residential\_High\_Density Residential\_Low\_Density   
## 20 1586   
## Residential\_Medium\_Density   
## 325

#Some neighborhoods have very few (only 1) home sale while others have many (327)

Ames\_gp\_neighborhoods <- ames %>%  
 group\_by(Neighborhood) %>%  
 summarize(freq = n()) %>%  
 arrange(desc(freq))

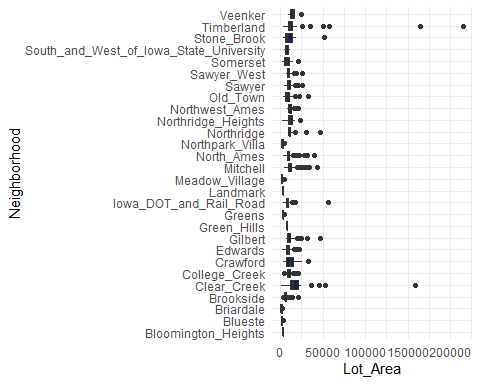
Ames\_gp\_neighborhoods

## # A tibble: 28 × 2  
## Neighborhood freq  
## <fct> <int>  
## 1 North\_Ames 326  
## 2 College\_Creek 183  
## 3 Old\_Town 180  
## 4 Edwards 127  
## 5 Somerset 118  
## 6 Gilbert 109  
## 7 Sawyer 109  
## 8 Northridge\_Heights 104  
## 9 Northwest\_Ames 95  
## 10 Sawyer\_West 82  
## # … with 18 more rows

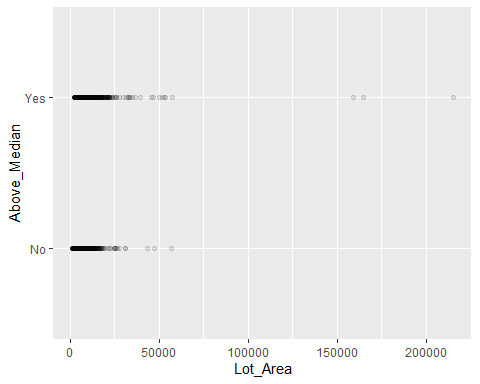
#Year Sold is an integer variable, but we want to examine each year as it’s own category. Therefore, we convert Year\_Sold into a factor.

ames\_yr\_as\_factor <- ames %>%  
 mutate(Year\_Sold = as\_factor(Year\_Sold))

ggplot(ames) +  
 aes(x = Lot\_Area, y = Neighborhood) +  
 geom\_boxplot(fill = "#112446") +  
 theme\_minimal()



ggplot(ames, aes(x=Lot\_Area, y=Above\_Median))+  
 geom\_point(alpha=.1)



ames <- ames %>%  
select(c(1:4,6:8,10:79))  
#Removing columns 5 & 9 (Street and Utilities) as these variables that don't add to our predictions.

#Remove Alley, Land\_Slope, Condition\_1, Condition\_2, Sale\_Condition, Sale\_type, Basement\_Type\_2\_Misc\_features as all have no predictive value