FINAL PROJECT - PHASE 1

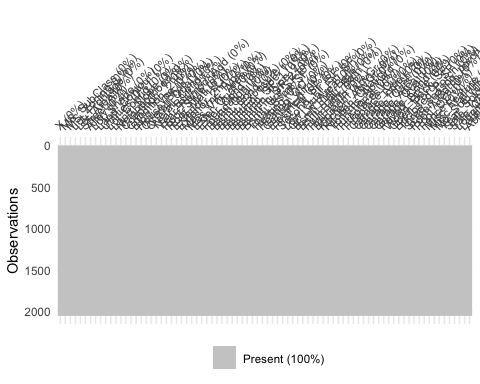
Charlie Wilds and Steven Booksh

3/1/2021

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

# CHECKING FOR MISSING DATA

vis\_miss(student)



## OBSERVATION AND VARIABLE TOTALS

nrow(student)

## [1] 2053

ncol(student)

## [1] 82

## STRUCTURE AND SUMMARY OF DATA SET

# str(student)  
# summary(student)

## RECODING CHARACTER VARIABLES INTO FACTORS

student = student %>%  
mutate(MS\_SubClass=as\_factor(MS\_SubClass), MS\_Zoning=as\_factor(MS\_Zoning), Street=as\_factor(Street), Alley=as\_factor(Alley), Lot\_Shape=as\_factor(Lot\_Shape), Land\_Contour=as\_factor(Land\_Contour), Utilities=as\_factor(Utilities), Lot\_Config=as\_factor(Lot\_Config), Land\_Slope = as\_factor(Land\_Slope), Neighborhood = as\_factor(Neighborhood), Condition\_1 = as\_factor(Condition\_1), Condition\_2 = as\_factor(Condition\_2), Bldg\_Type = as\_factor(Bldg\_Type), House\_Style = as\_factor(House\_Style), Overall\_Qual = as\_factor(Overall\_Qual), Overall\_Cond = as\_factor(Overall\_Cond), Roof\_Style = as\_factor(Roof\_Style), Roof\_Matl = as\_factor(Roof\_Matl), Exterior\_1st = as\_factor(Exterior\_1st), Exterior\_2nd = as\_factor(Exterior\_2nd), Mas\_Vnr\_Type = as\_factor(Mas\_Vnr\_Type), Mas\_Vnr\_Area = as\_factor(Mas\_Vnr\_Area), Exter\_Qual = as\_factor(Exter\_Qual), Exter\_Cond = as\_factor(Exter\_Cond), Foundation = as\_factor(Foundation), Bsmt\_Qual = as\_factor(Bsmt\_Qual), Bsmt\_Cond =as\_factor(Bsmt\_Cond), Bsmt\_Exposure = as\_factor(Bsmt\_Exposure), BsmtFin\_Type\_1 = as\_factor(BsmtFin\_Type\_1), BsmtFin\_Type\_2 = as\_factor(BsmtFin\_Type\_2), Heating =as\_factor(Heating), Heating\_QC = as\_factor(Heating\_QC), Central\_Air = as\_factor(Central\_Air), Electrical = as\_factor(Electrical), Kitchen\_Qual = as\_factor(Kitchen\_Qual), Functional = as\_factor(Functional), Fireplace\_Qu = as\_factor(Fireplace\_Qu), Garage\_Type = as\_factor(Garage\_Type), Garage\_Finish = as\_factor(Garage\_Finish), Garage\_Qual = as\_factor(Garage\_Qual), Garage\_Cond = as\_factor(Garage\_Cond), Paved\_Drive =as\_factor(Paved\_Drive), Pool\_QC = as\_factor(Pool\_QC), Fence = as\_factor(Fence), Misc\_Feature = as\_factor(Misc\_Feature), Sale\_Type = as\_factor(Sale\_Type), Sale\_Condition = as\_factor(Sale\_Condition))

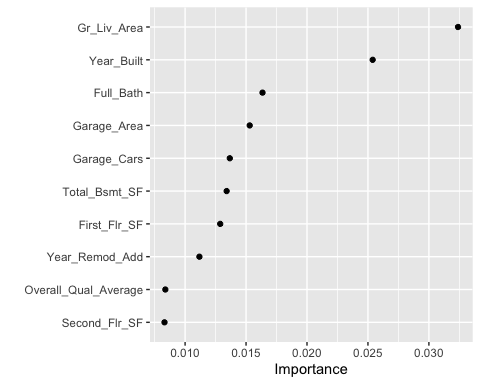
## RANDOM FOREST MODEL

rf\_recipe = recipe(Above\_Median ~., student) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_model = rand\_forest() %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
rf\_wflow =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(rf\_recipe)  
  
set.seed(123)  
rf\_fit = fit(rf\_wflow, student)  
  
rf\_fit

## ══ Workflow [trained] ══════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 1 Recipe Step  
##   
## ● step\_dummy()  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
## Ranger result  
##   
## Call:  
## ranger::ranger(x = maybe\_data\_frame(x), y = y, importance = ~"permutation", num.threads = 1, verbose = FALSE, seed = sample.int(10^5, 1), probability = TRUE)   
##   
## Type: Probability estimation   
## Number of trees: 500   
## Sample size: 2053   
## Number of independent variables: 682   
## Mtry: 26   
## Target node size: 10   
## Variable importance mode: permutation   
## Splitrule: gini   
## OOB prediction error (Brier s.): 0.0614444

## **VARIABLE IMPORTANCE**

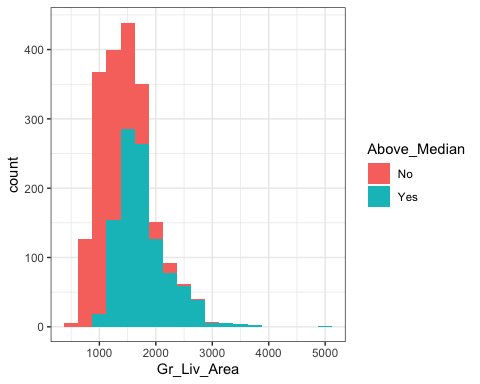
rf\_fit %>% pull\_workflow\_fit() %>% vip(geom = "point")



# **STRONG PREDICTORS OF Above\_Median HOME SALES VARIABLE**

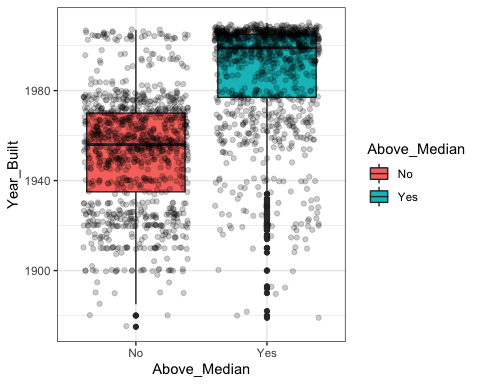
### *VISUALIZATION FOR Ground Living Area (Gr\_Liv\_Area)*

ggplot(student, aes(x=Gr\_Liv\_Area,fill=Above\_Median)) + geom\_histogram(binwidth=250) + theme\_bw()



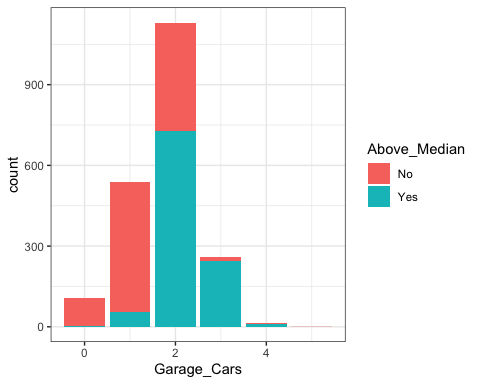
### *VISUALIZATION FOR Year Built (Year\_Built)*

ggplot(student, aes(x=Above\_Median,y=Year\_Built)) + geom\_boxplot(aes(fill=Above\_Median)) + theme\_bw() + geom\_jitter(alpha=.2)



### *VISUALIZATION FOR Garage Cars (Garage\_Cars)*

ggplot(student, aes(x=Garage\_Cars,fill=Above\_Median)) + geom\_bar() + theme\_bw()



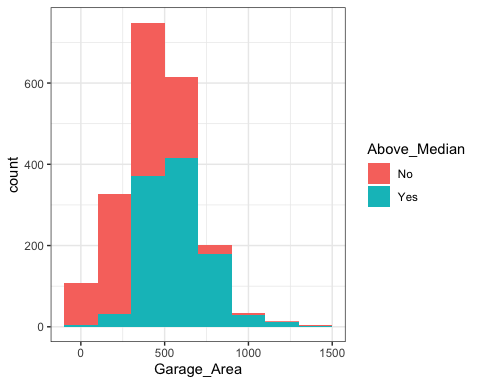
### *ADDITIONAL TABLE FOR Garage Cars (Garage\_Cars)*

t1 = table(student$Above\_Median, student$Garage\_Cars)  
prop.table(t1, margin = 2)

##   
## 0 1 2 3 4 5  
## No 0.96296296 0.89610390 0.35632184 0.06130268 0.23076923 1.00000000  
## Yes 0.03703704 0.10389610 0.64367816 0.93869732 0.76923077 0.00000000

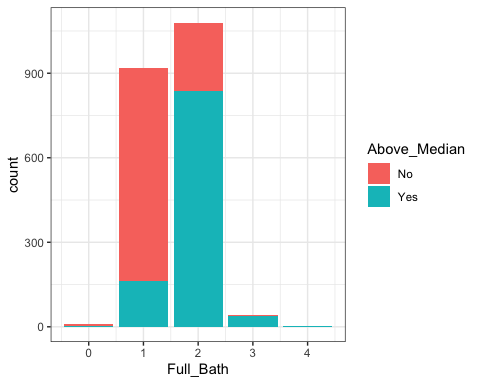
### *VISUALIZATION FOR Garage Area (Garage\_Area)*

ggplot(student, aes(x=Garage\_Area,fill=Above\_Median)) + geom\_histogram(binwidth=200) + theme\_bw()



### *VISUALIZATION FOR Full Bathrooms (Full\_Bath)*

ggplot(student, aes(x=Full\_Bath,fill=Above\_Median)) + geom\_bar() + theme\_bw()



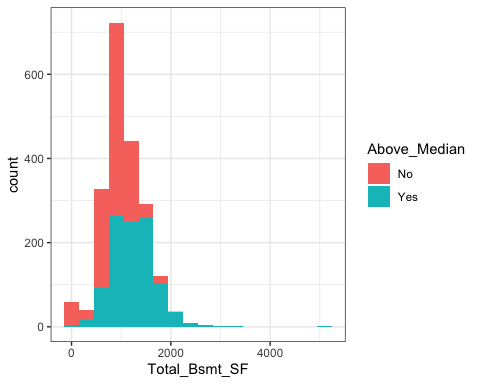
### *ADDITIONAL TABLE FOR Full Bathrooms (Full\_Bath)*

t1 = table(student$Above\_Median, student$Full\_Bath)  
prop.table(t1, 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

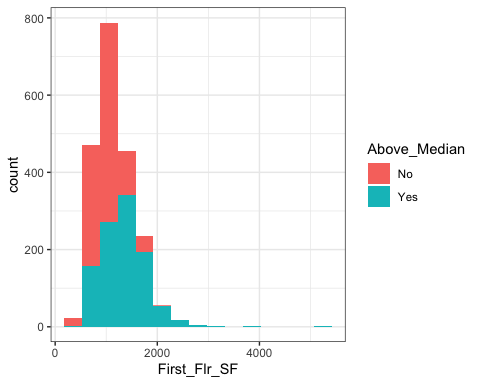
### *VISUALIZATION FOR Total Basement Square Footage (Total\_Bsmt\_SF)*

ggplot(student, aes(x=Total\_Bsmt\_SF,fill=Above\_Median)) + geom\_histogram(binwidth=300) + theme\_bw()



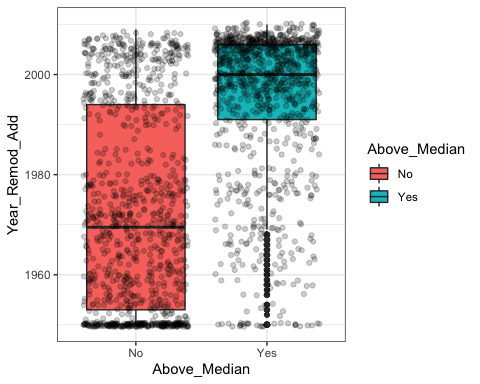
### *VISUALIZATION FOR First Floor Square Footage (First\_Flr\_SF)*

ggplot(student, aes(x=First\_Flr\_SF,fill=Above\_Median)) + geom\_histogram(binwidth=350) + theme\_bw()



### *VISUALIZATION FOR Year Home was Remodeled (Year\_Remod\_Add)*

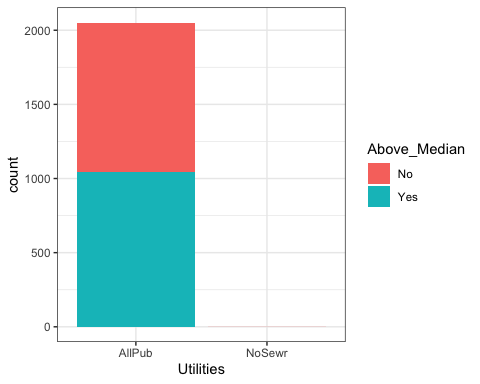
ggplot(student, aes(x=Above\_Median,y=Year\_Remod\_Add)) + geom\_boxplot(aes(fill=Above\_Median)) + theme\_bw() + geom\_jitter(alpha=0.2)



## **WEAK PREDICTORS OF Above\_Median HOME SALES VARIABLE**

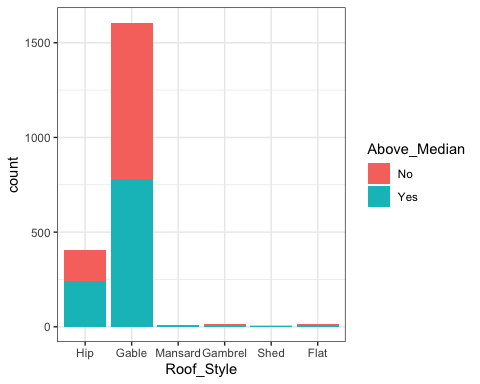
### *VISUALIZATION FOR Utilities (Utilities)*

ggplot(student, aes(x=Utilities, fill=Above\_Median)) + geom\_bar() + theme\_bw()



### *VISUALIZATION FOR Roof Style (Roof\_Style)*

ggplot(student, aes(x=Roof\_Style,fill=Above\_Median)) + geom\_bar() + theme\_bw()



### *VISUALIZATION FOR Street Type (Street)*

ggplot(student, aes(x=Street, fill=Above\_Median)) + geom\_bar() + theme\_bw()

