Final Project Phase 1

Christopher Dillard

2023-06-18

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

ames = read\_csv("ames\_student-1.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)  
unique(ames$Year\_Sold)

## [1] 2010 2009 2008 2007 2006

1. Tidying Data and General Prep/New Variable Creation and Selection:

any(is.na(ames))

## [1] FALSE

Luckily, this data set contains no NA values.

The variables I am interested in exploring are:

1. Neighborhood
2. Lot\_Area
3. Year\_Built and Year\_Remod\_Add, if applicable
4. Overall\_Cond (Condition)
5. MS\_Zoning (primarily to compare low-density and high-density later in Phase 2 of the Project)

While I explore these variables and their potential significance in predicting whether a home will be sold above the Median price, I want to consider demographic information, in particular the socio-economic factors that could influence pricing, in addition to some of the variables included in the dataset. My “quick facts” in the power point are taken from recent [US Census data for Ames,Iowa](https://www.census.gov/quickfacts/amescityiowa).

So, I’ll create two new columns that can be of use:

1. “Remodeled”, which has 2 levels for either “Remodeled” or “Not Remodeled” based on whether the values in the Year\_Remod\_Add column are greater than their corresponding values in Year\_Built.

ames <- ames %>%  
 mutate(  
 Remodeled = ifelse(Year\_Remod\_Add > Year\_Built, "Remodeled", "Not Remodeled"),  
 Remodeled = factor(Remodeled)  
 )

1. “Decade”, which groups the Year\_Built values into 10 year increments. This is really just to make the graphs easier to read.

#first, what year what the oldest house sold built?  
min(ames$Year\_Built)

## [1] 1875

So, we start our decades at 1870:

ames <- ames %>%  
 mutate(  
 Decade = floor(Year\_Built / 10) \* 10) #AKA we drop the last digit in the year and replace it with a zero

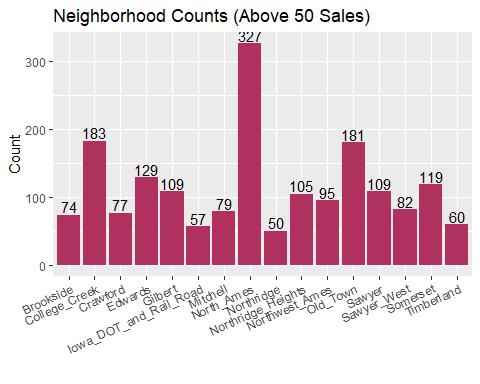
Which Neighborhoods are shown in the dataset?

unique(ames$Neighborhood)

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

The following graph will display only the neighborhoods with at least 50 sales, so that it is easier to read and straight to the point.

neighborhoodcount2 = ames %>%   
 count(Neighborhood) %>%  
 top\_n(16,n)  
  
ggplot(neighborhoodcount2, aes(x = Neighborhood, y = n)) +  
 geom\_bar(stat="identity", fill = "maroon") +  
 geom\_text(aes(label = n), vjust = -.2)+  
 xlab("") +  
 ylab("Count") +  
 ggtitle("Neighborhood Counts (Above 50 Sales)") +  
 theme(axis.text.x = element\_text(angle = 25, hjust = 1))



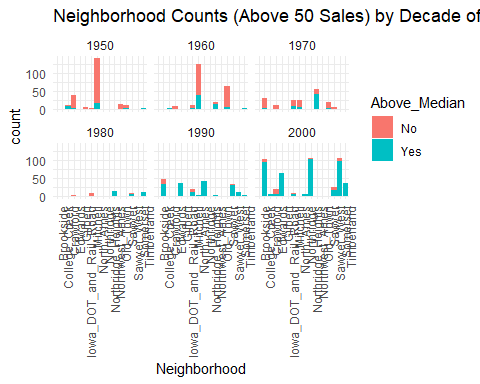
I’m convinced that location is an important feature in the determining the sales price of a home so it would be interesting to compare the prices for the Neighborhoods with the highest (chiefly North Ames) with those that registered relatively few sales and did not meet the minimum cut-off of 50 sales.

Time to look at homes sold above median sales price in each neighborhood and facet by the decade they were built in (over 1950):

#First I make a new revised dataset with the variables I'm interst in using.   
ames\_revised = ames %>%  
 select(Above\_Median, Remodeled, Year\_Built, Year\_Remod\_Add, Decade, Neighborhood, MS\_Zoning,Lot\_Area)  
#Reference the neighborhoodcount2 object I made earlier I filter for just the top 16 Neighborhoods, and store back into the ames\_revised df.  
ames\_revised <- ames\_revised %>%  
 filter(Neighborhood %in% neighborhoodcount2$Neighborhood)

The following graph shows that the order of the Neighborhoods along the x-axis is consistent with the order seen in the previous graph for comparative purposes. This graph can potentially tells us which neighborhoods are older, some insight into when most of the housing construction happened in Ames, and perhaps which decades saw an improvement in construction standards that could influence selling price. These are not 100% compelling though since this data only includes housing being sold in each neighborhood, which could be misleading when looking into the volume of sales in neighborhoods dominated by off-campus housing, which is arguably higher than older residential neighborhoods in cities with Universities. The neighborhood names will be dropped for this particular graph in the powerpoint since they were always displayed in the previous slide.

ames\_revised %>%  
 mutate(Decade = as.numeric(Decade)) %>%  
 filter(Decade >= 1950 & Decade <= 2000) %>%  
 ggplot() +  
 aes(x = Neighborhood, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 theme\_minimal() +  
 facet\_wrap(vars(Decade)) +  
 ggtitle("Neighborhood Counts (Above 50 Sales) by Decade of Home's Construction") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



I think lot area, by square footage in the dataset with the “Lot\_Area” variable, plays another important role in determining sales price.

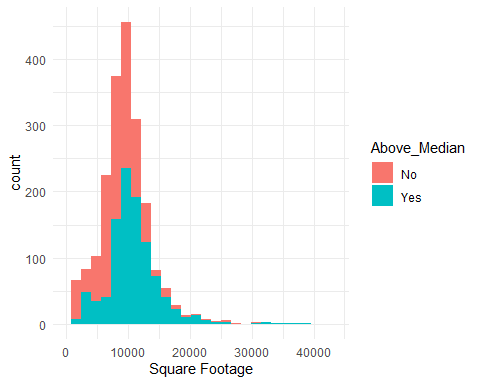
ames %>%   
 group\_by(Neighborhood) %>%  
 summarize(Avg = mean(Lot\_Area),Min=min(Lot\_Area),Max=max(Lot\_Area))

## # A tibble: 28 × 4  
## Neighborhood Avg Min Max  
## <chr> <dbl> <dbl> <dbl>  
## 1 Bloomington\_Heights 3350. 3010 3940  
## 2 Blueste 2042 1300 3907  
## 3 Briardale 1816. 1680 2368  
## 4 Brookside 7081. 3500 21384  
## 5 Clear\_Creek 22998. 1700 159000  
## 6 College\_Creek 10014. 4426 21533  
## 7 Crawford 11867. 3842 32668  
## 8 Edwards 9940. 2522 47007  
## 9 Gilbert 10981. 7250 47280  
## 10 Green\_Hills 8239 8239 8239  
## # ℹ 18 more rows

ames %>%  
 filter(!(Roof\_Matl %in% "Tar&Grv")) %>%  
 ggplot() +  
 aes(x = Lot\_Area, fill = Above\_Median) +  
 geom\_histogram(bins = 28L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(  
 x = "Square Footage")+  
 theme\_minimal() +  
 xlim(0, 43560)

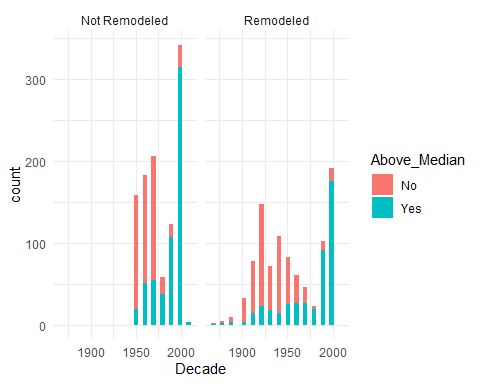
## Warning: Removed 12 rows containing non-finite values (`stat\_bin()`).

## Warning: Removed 4 rows containing missing values (`geom\_bar()`).



The graph below does not show any compelling evidence that remodels houses sell above Median Price. It also does not make the distinction between WHEN a house was remodeled, which should be relevant. So, I’ll be omitting the Remodeled variable from my power point, but won’t entirely discard it from my model in Phase 2.

ames %>%  
 filter(!(Roof\_Matl %in% "Tar&Grv")) %>%  
 ggplot() +  
 aes(x = Decade, fill = Above\_Median) +  
 geom\_histogram(bins = 30L) +  
 scale\_fill\_hue(direction = 1) +  
 theme\_minimal() +  
 facet\_wrap(vars(Remodeled))

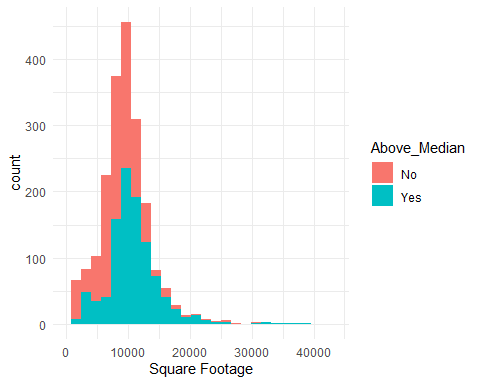


Now, let’s look at how a home’s square footage can influence whether they sold above Median price or not:

ames %>%  
 filter(!(Roof\_Matl %in% "Tar&Grv")) %>%  
 ggplot() +  
 aes(x = Lot\_Area, fill = Above\_Median) +  
 geom\_histogram(bins = 28L) +  
 scale\_fill\_hue(direction = 1) +  
 labs(  
 x = "Square Footage")+  
 theme\_minimal() +  
 xlim(0, 43560)

## Warning: Removed 12 rows containing non-finite values (`stat\_bin()`).

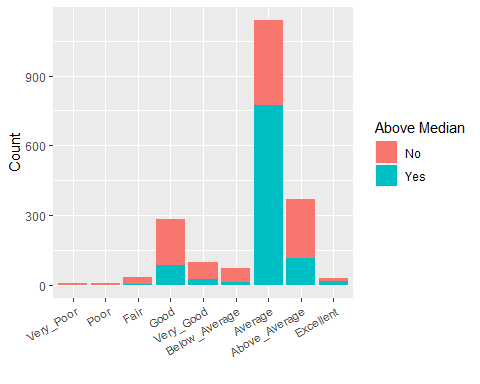
## Warning: Removed 4 rows containing missing values (`geom\_bar()`).



To Plot Overall Condition, there are a few options at hand:

Let’s simply look at how the proportion of homes sold above median price for each Overall Condition.

ames %>%  
 filter(!(Roof\_Matl %in% "Tar&Grv")) %>%  
 ggplot() +  
 aes(x = factor(Overall\_Cond, levels = c("Very\_Poor", "Poor", "Fair", "Good", "Very\_Good", "Below\_Average", "Average", "Above\_Average", "Excellent")), fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 labs(x="",y="Count",fill="Above Median")+  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))



Perhaps, how homes compare within the same neighborhood depending on whether they have been remodeled or not?

ggplot(ames\_revised) +  
 aes(x = Neighborhood, fill = Above\_Median) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 theme\_minimal() +  
 facet\_wrap(vars(Remodeled))+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

