Analyzing Ames Housemarket

Brandon Leff

4/2/2019

```
knitr::opts_chunk$set(echo = TRUE)
ames <- read.csv("/Users/brandonleff/Desktop/math130/data/ames.csv", header=TRUE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(ggplot2)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(scales)
```

Introduction

We will be analyzing the `ames` data set which is a dataset of all residential home sales in Ames, Iowa between 2006 and 2010. We will be taking a look into a few variables including, `Year.Remod.Add` which is the year in which a remodling was done to the house, `Year.Built` which is the year the house was built, `Roof.Style` which is the type of roof style for each house, and lastly, `SalePrice` which is the sale price of the house when it was sold between 2006 and 2010.

Analysis of House Prices in Ames

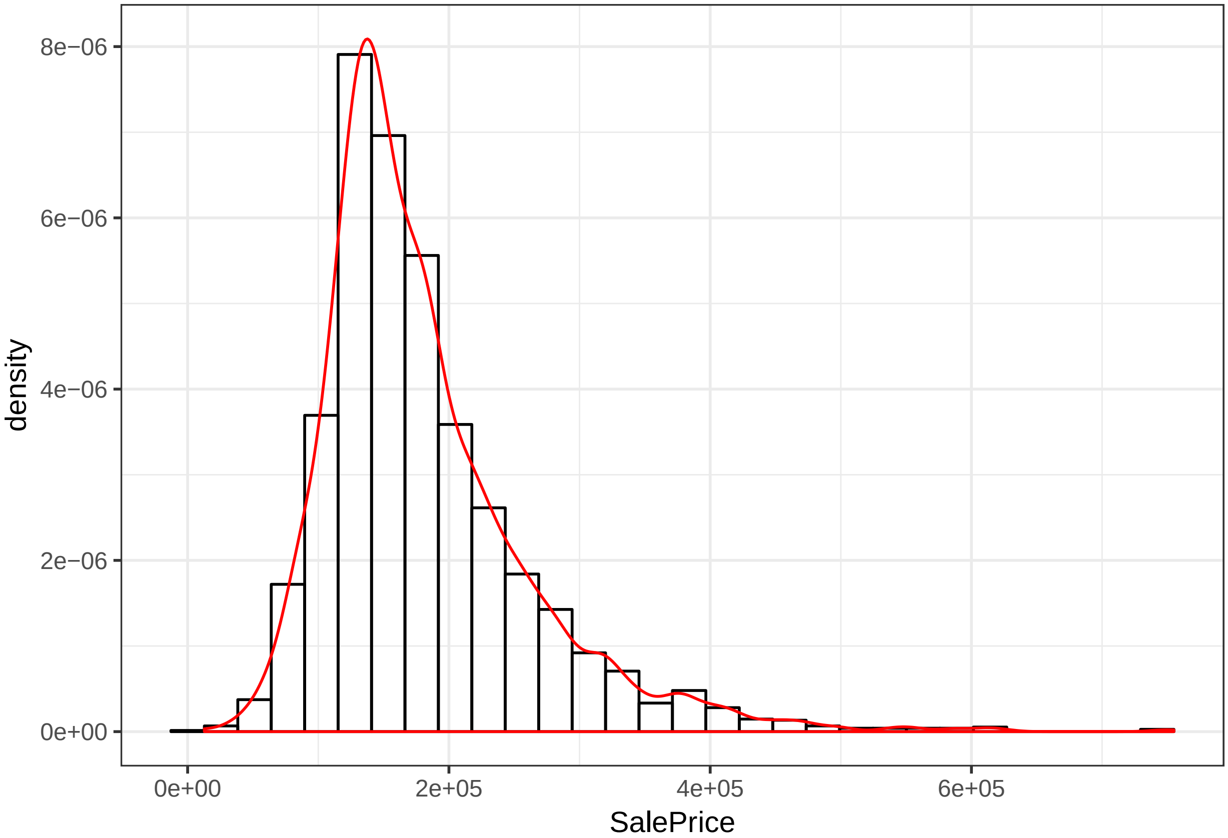
Our first task is that we want to see the spread and summary statistics of the house prices in Ames during this 4 year period. We want to see if it is normally distributed or skewed in a certain direction which can give us insight on the status of the city.

```
summary(ames$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12789  129500  160000  180796  213500  755000
```

```
ggplot(ames, aes(x=SalePrice)) + geom_histogram(aes(y = ..density..), col = "black", fill = NA) + theme_
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram of `SalePrice` is skewed right with a higher amount of houses around the `$100,000 - $200,000` price range. There are more expensive houses that both pull the mean higher and also skew the graph right.

Relationship Between Recent Renovations and House Price

First, we will create a new variable `between.ren` which is the amount of years in between the latest renovation and the year the house was originally built. Then we will confirm that the new variable was made correctly.

```
ames$between.ren <- ames$Year.Remod.Add - ames$Year.Built
head(ames[c("Year.Remod.Add", "Year.Built", "between.ren")])
```

```
##   Year.Remod.Add Year.Built between.ren
## 1           1960       1960           0
## 2           1961       1961           0
## 3           1958       1958           0
## 4           1968       1968           0
## 5           1998       1997           1
## 6           1998       1998           0
```

After building this new variable, we don't want to look at the houses that have never been renovated or have been entered in incorrectly, so we will remove the values less than or equal to `0`.

```
ames$between.ren[ames$between.ren<=0] <- NA
summary(ames$between.ren)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00    1.00   22.00   27.81   44.00  127.00    1570
```

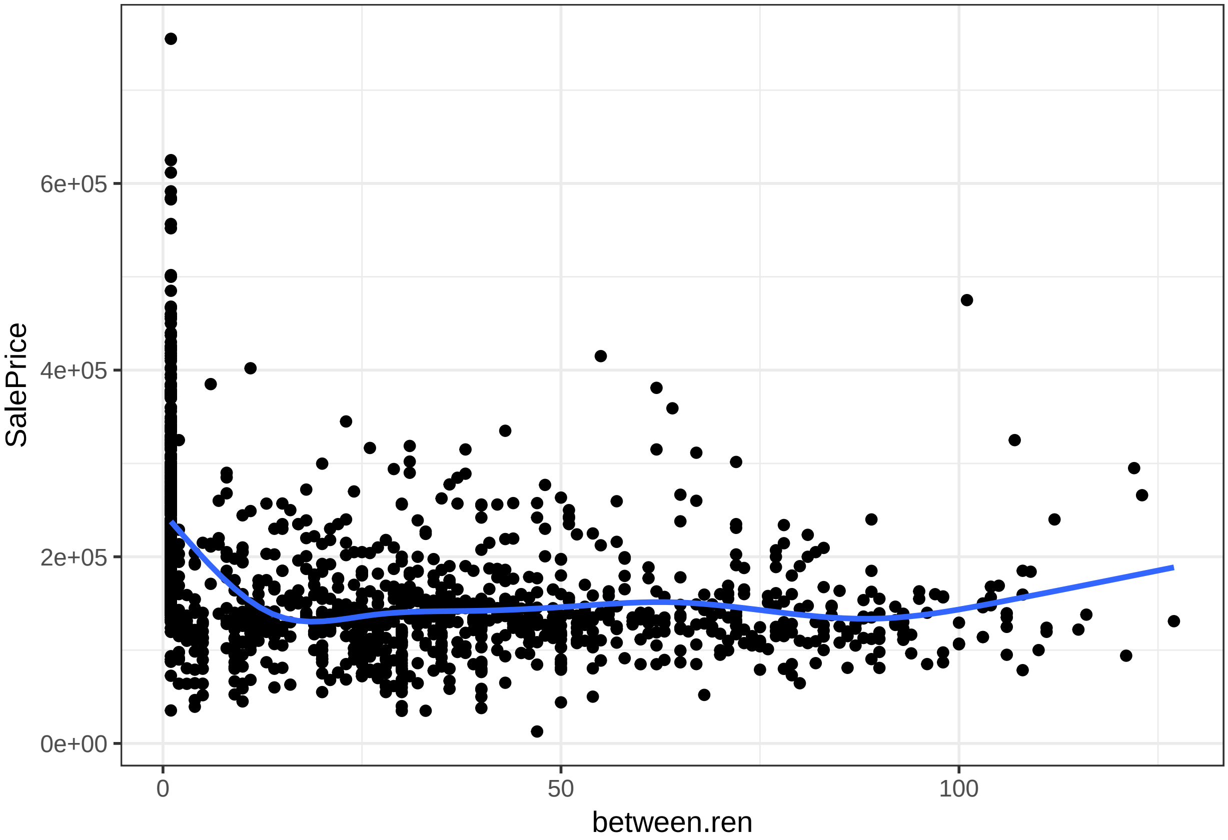
This new variable is showing the amount of years between the house being built and the latest renovation, so a house with a high number in `between.ren` would represent a house built a long time ago with a relatively recent renovation.

```
ggplot(ames, aes(x=between.ren, y=SalePrice)) + geom_point() + geom_smooth(se=FALSE) + theme_bw()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1570 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1570 rows containing missing values (geom_point).
```



With a combination of the first quartile of data being at `1` and also the factor that `between.ren` doesn't take count of how recent the renovation is relative to the current year, there seems to be no defined relationship between the amount of years between renovation and the house price.

Relationship Between Roof Type and Year Built

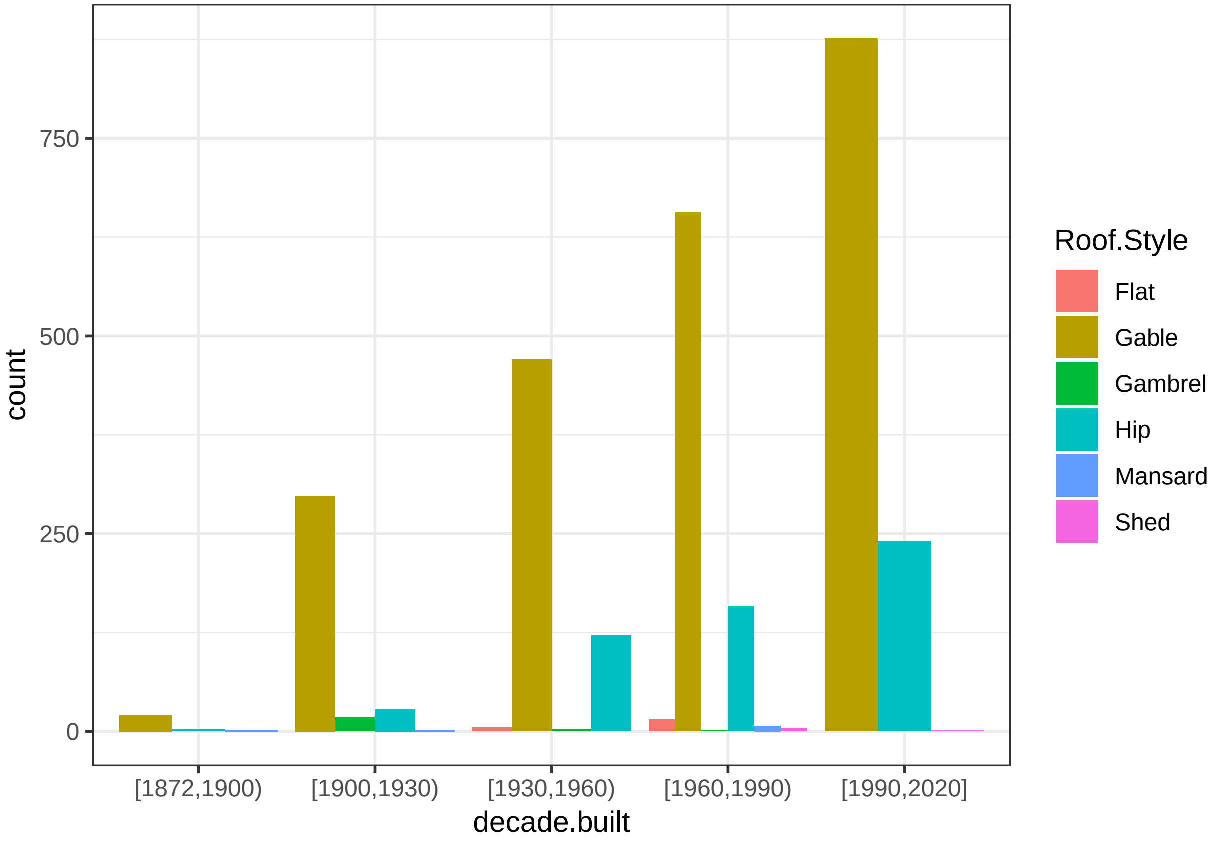
We are now interested in the architecture of the region. We want to know the trends of roof types and if certain time periods had a preference in the roof type. First we have to take the variable `Year.Built` and scale it into decades and make that a new variable `decade.built`.

```
ames$decade.built <- cut2(ames$Year.Built, c(1900, 1930, 1960, 1990, 2020))
table(ames$decade.built)
```

```
##
## [1872,1900) [1900,1930) [1930,1960) [1960,1990) [1990,2020]
##          26         346         600         841        1117
```

We broke up the data in `Year.Built` into 5 30 year periods starting from 1872 and ending at 2020. Although no houses were built in 2020, the interval still encaptures the house built the latest in this data set without any harm in the other intervals. Next we will create a group bar chart comparing the roof types in each decade chunk to see if there are patterns.

```
ggplot(ames, aes(x=decade.built, fill=Roof.Style)) + geom_bar(position = "dodge") + theme_bw()
```



The graph shows that predominatly the Gable roof was popular throughout all the decade chunks being the roof style with the biggest count for all decade chunks. The Hip style roof originally started gaining popularity in the `[1900,1930)` period and then continued being the second most popular roof style for the rest of the time. Gambrel roof styles had slight popularity also in the `[1900,1930)` time period but actually died down in the later years. Finally, Flat roof styles were mostly seen in the `[1960,1990)` time period.