Stat 480 - Homework #6

Vahid Azizi

2/26/2020

Ames housing

1. Download the RMarkdown file with these homework instructions to use as a template for your work. Make sure to replace “Your Name” in the YAML with your name.
2. The Ames based, non-profit company OAITI provides aoe open-source data sets. One of these data sets consists of information on all house sales in Ames between 2008 and 2010. The following piece of code allows you to read the dataset into your R session. How many house sales were there between 2008 and 2010? Which type of variables are we dealing with?

```
housing <- read.csv("https://raw.githubusercontent.com/OAITI/open-datasets/master/Housing%20Dat
a/Ames-Housing.csv")
```

```
paste("Number of house sales between 2008 to 2010 is:", nrow(housing))
```

```
## [1] "Number of house sales between 2008 to 2010 is: 1615"
```

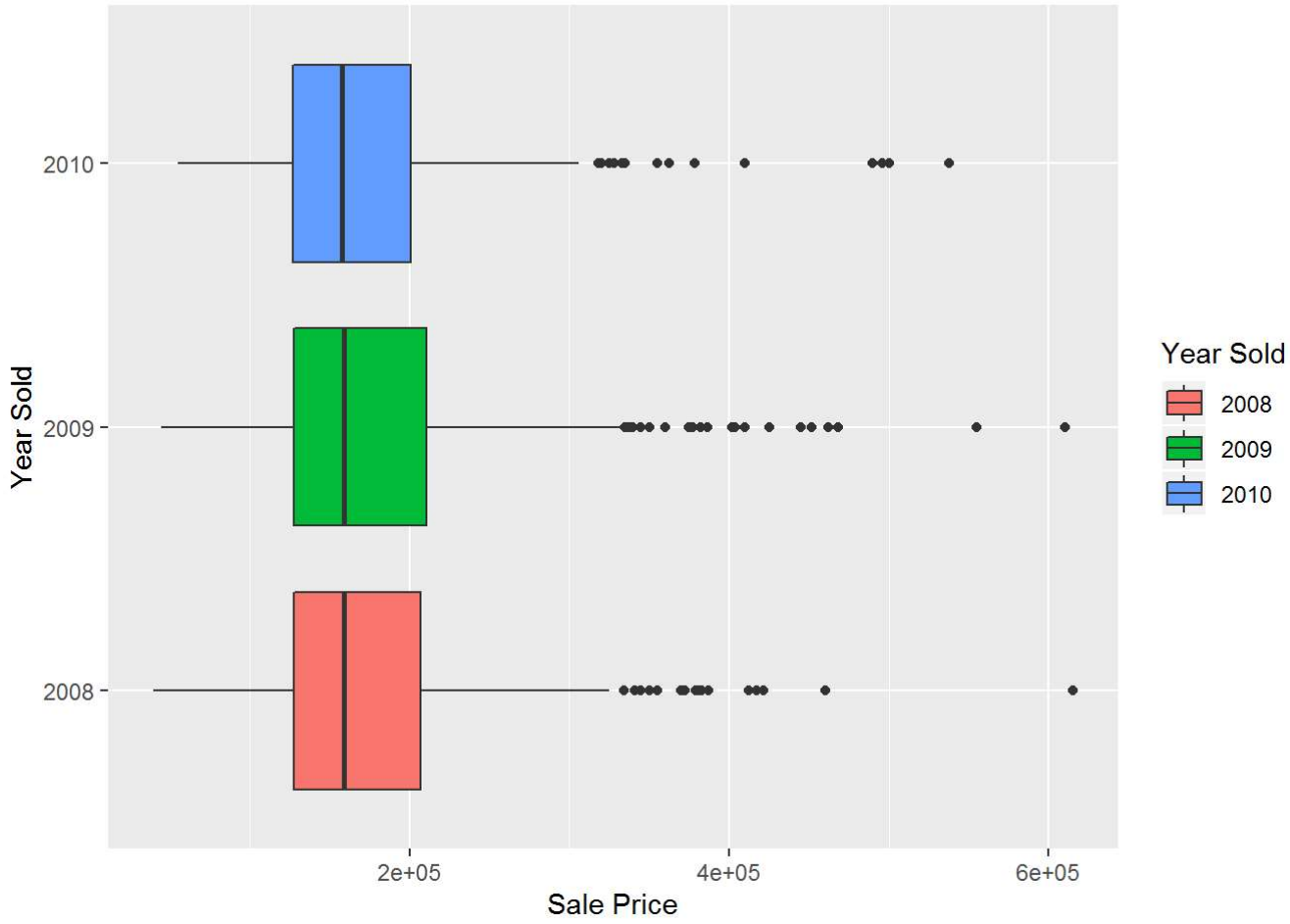
```
str(housing)
```

```
## 'data.frame':    1615 obs. of  10 variables:
##  $ SalePrice   : int  215000 124500 105000 172000 176500 157000 244000 237500 206900 345000
...
##  $ Bedrooms    : int  3 2 2 3 3 4 3 4 4 4 ...
##  $ Baths       : int  1 1 1 1 1 2 2 2 2 2 ...
##  $ LotArea     : int  31770 13008 11622 14267 11029 10200 11160 12925 11075 13860 ...
##  $ LivingArea  : int  1656 882 896 1329 1414 1434 2110 2117 2112 2704 ...
##  $ GarageArea  : int  528 502 730 312 601 528 522 550 576 538 ...
##  $ Neighborhood: Factor w/ 33 levels "Blmngtn","Blueste",..: 19 19 19 19 19 19 19 19 19 19
...
##  $ HouseStyle  : Factor w/ 8 levels "1-Story","1.5 Fin",..: 1 1 1 1 1 1 1 1 4 8 ...
##  $ YearBuilt   : int  1960 1956 1961 1958 1958 1974 1968 1970 1969 1972 ...
##  $ YearSold    : int  2010 2009 2010 2010 2008 2009 2010 2008 2008 2009 ...
```

Two out of 10 variables are factor and the rest of them are int.

3. Do sales prices change over time? (Don’t test significances) Provide a graphic that supports your statement.

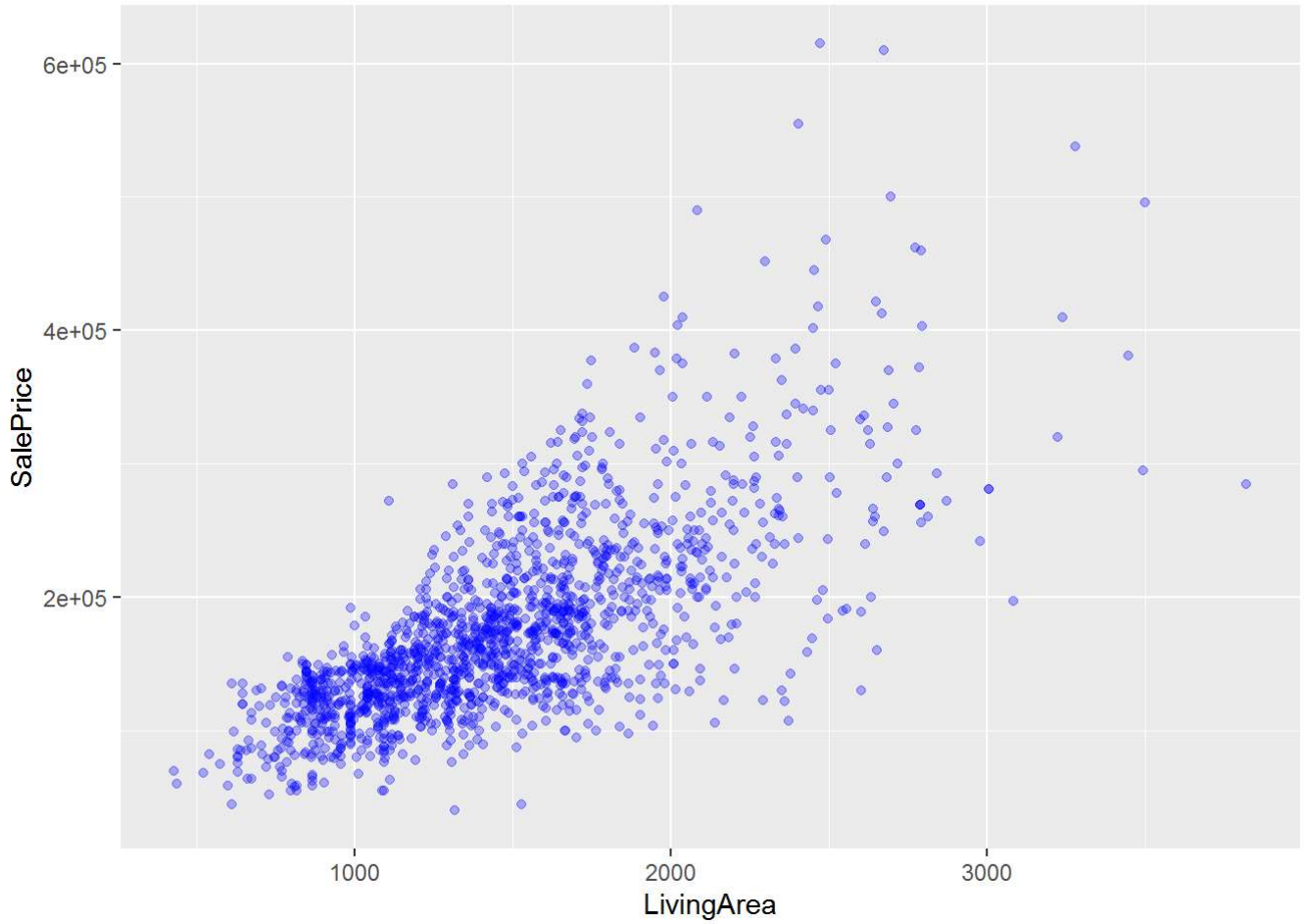
```
ggplot(housing,aes(x=factor(housing$YearSold) ,y=housing$SalePrice, fill=factor(YearSold)))+
  geom_boxplot()+
  coord_flip()+
  labs(x ="Year Sold", y = "Sale Price", fill="Year Sold")
```



The prices have not been changed significantly on average, but by taking a precise look at boxplots, it seems that in 2010 sale price is decreased a little bit because the third quartile of this box is less than the other boxplots' third quartile. Also, it seem box plot for year 2010 has less number of outliers.

4. What is the relationship between sales prices and the size of the house (living area)? Make a chart and describe the relationship.

```
ggplot(housing,aes(y=SalePrice,x=LivingArea))+
  geom_point(color = "blue", alpha = .3)
```



By increasing “Living Area” house price increases.

5. Use `dplyr` functions to:

- introduce a variable consisting of price per square foot,
- find the average price per square foot in each of the Ames neighborhoods,
- exclude averages that are based on fewer than 10 records,
- reorder the remaining neighborhoods according to the mean sales prices.

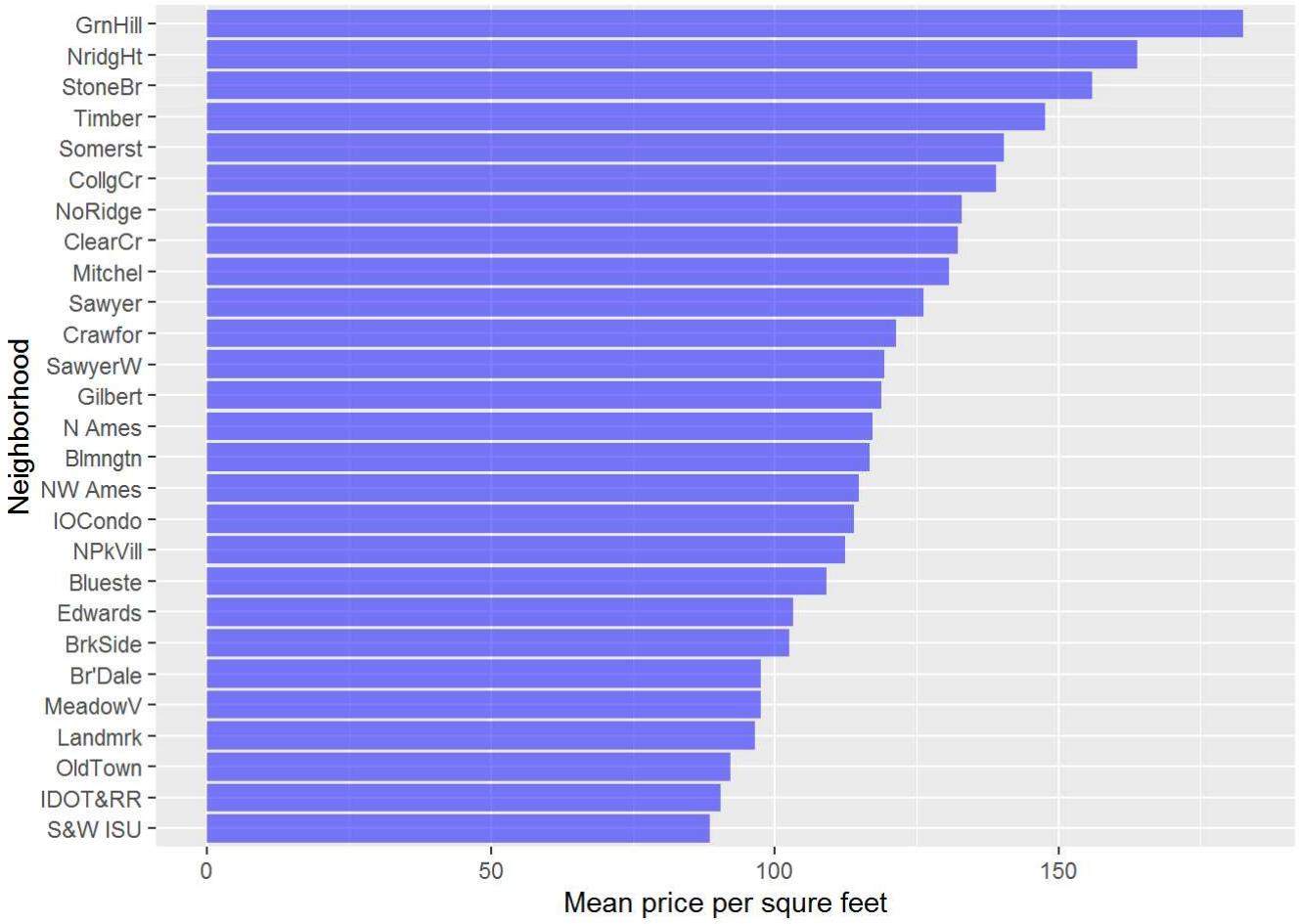
```
housing %>% mutate(ppsf=housing$SalePrice/housing$LivingArea) %>% group_by(Neighborhood) %>% sum
marise(mean_ppsf=mean(ppsf,na.rm=TRUE),number_records=length(ppsf)) %>% filter(number_records>9)
%>% arrange(mean_ppsf)
```

```
## # A tibble: 27 x 3
##    Neighborhood mean_ppsf number_records
##    <fct>            <dbl>          <int>
##  1 S&W ISU           88.6             30
##  2 IDOT&RR           90.5             39
##  3 OldTown           92.2            127
##  4 Landmrk           96.5             20
##  5 MeadowV           97.5             21
##  6 Br'Dale           97.5             16
##  7 BrkSide          103.              63
##  8 Edwards          103.              95
##  9 Blueste          109.              19
## 10 NPkVill          112.              17
## # ... with 17 more rows
```

6. Draw a chart of the average sale prices by neighborhood and comment on it. Only consider neighborhoods with at least 10 sales.

   Bonus: write the code for this question and the previous one in a single statement for +0.5 point extra credit.

```
housing %>% mutate(ppsf=housing$SalePrice/housing$LivingArea) %>% group_by(Neighborhood) %>% sum
marise(mean_ppsf=mean(ppsf,na.rm=TRUE),number_records=length(ppsf)) %>% filter(number_records>9)
%>% arrange(mean_ppsf) %>%  ggplot(aes(x=fct_reorder(Neighborhood,mean_ppsf),y=mean_ppsf))+
  geom_bar(stat = "identity",fill = "blue", alpha = .5)+
  coord_flip()+
  xlab("Neighborhood")+
  ylab("Mean price per squre feet")
```



7. Use `dplyr` functions to:

- introduce a logical variable called 'garage' that is FALSE if the garage area is zero, and TRUE otherwise,
- exclude all sales of houses that do not have a garage,
- only consider 1 and 2 story houses (`HouseStyle`),
- create a new variable `YBCut` from `YearBuilt` that introduces age categories that groups the year a house was built into intervals: 1800-1850, 1850-1900, 1950-2000, 2000+ (see `?cut`).

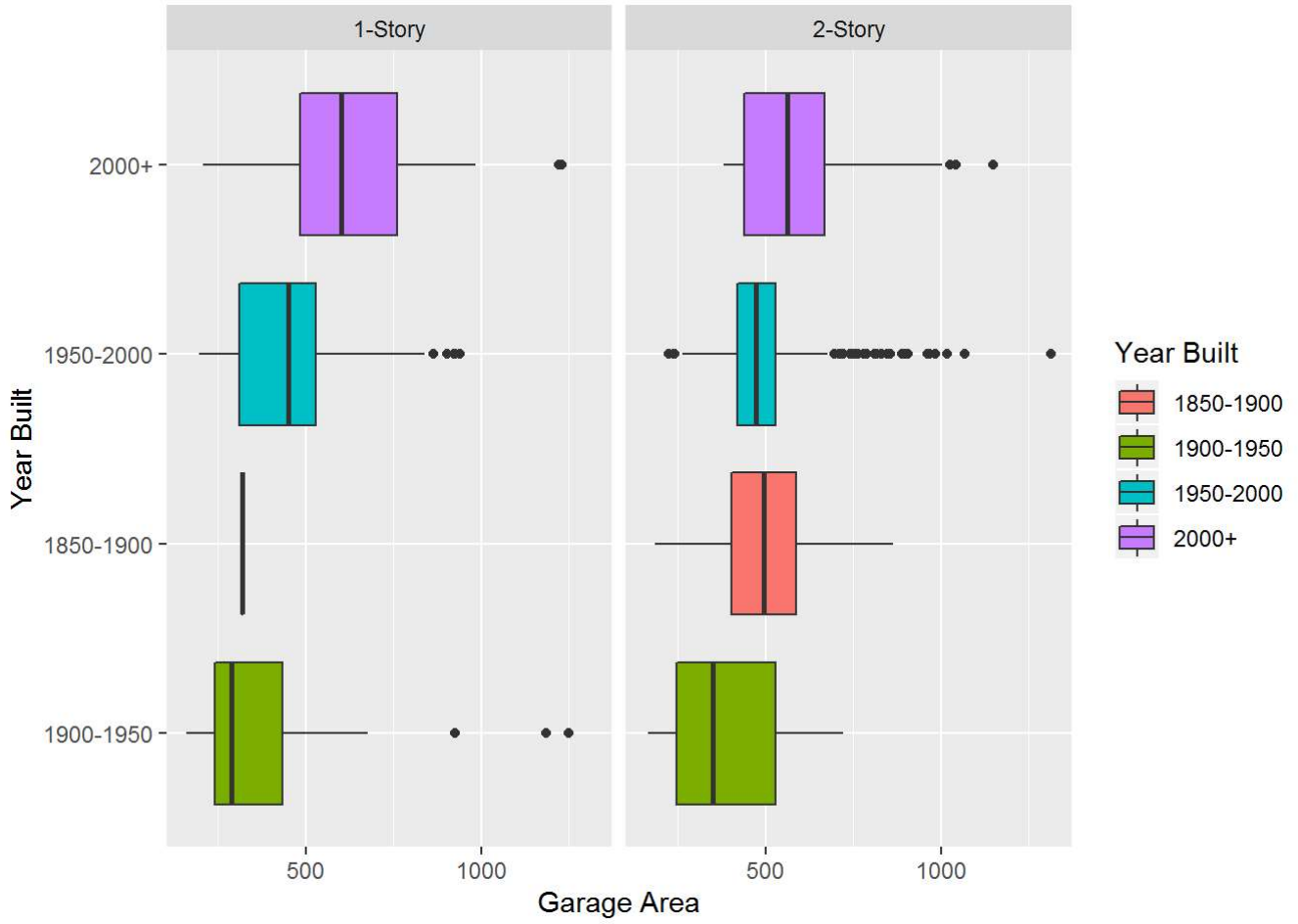
```
housing %>% mutate(garage=as.logical(case_when(GarageArea==0 ~ FALSE, GarageArea>0 ~ TRUE ))) %
>% filter(garage==TRUE) %>% filter(HouseStyle %in% c("1-Story","2-Story")) %>% mutate(YBCut=as.f
actor(case_when(YearBuilt>=1800 & YearBuilt<1850 ~ "1800-1850",
YearBuilt>=1850 & YearBuilt<1900 ~ "1850-1900",
YearBuilt>=1900 & YearBuilt<1950 ~ "1900-1950",
YearBuilt>=1950 & YearBuilt<2000 ~ "1950-2000",YearBuilt>=2000  ~ "2000+"))) %>% head()
```

```
##   SalePrice Bedrooms Baths LotArea LivingArea GarageArea Neighborhood
## 1    215000        3     1   31770       1656        528       N Ames
## 2    124500        2     1   13008        882        502       N Ames
## 3    105000        2     1   11622        896        730       N Ames
## 4    172000        3     1   14267       1329        312       N Ames
## 5    176500        3     1   11029       1414        601       N Ames
## 6    157000        4     2   10200       1434        528       N Ames
##   HouseStyle YearBuilt YearSold garage     YBCut
## 1    1-Story      1960     2010   TRUE 1950-2000
## 2    1-Story      1956     2009   TRUE 1950-2000
## 3    1-Story      1961     2010   TRUE 1950-2000
## 4    1-Story      1958     2010   TRUE 1950-2000
## 5    1-Story      1958     2008   TRUE 1950-2000
## 6    1-Story      1974     2009   TRUE 1950-2000
```

8. Draw a chart of the previous data set. Draw side-by-side boxplots of the garage area by `YBCut`. Facet by the style of house. Describe and summarise the chart.

   Bonus: write the code for this question and the previous one in a single statement for +0.5 point extra credit.

```
housing %>% mutate(garage=as.logical(case_when(GarageArea==0 ~ FALSE, GarageArea>0 ~ TRUE ))) %
>% filter(garage==TRUE) %>% filter(HouseStyle %in% c("1-Story","2-Story")) %>% mutate(YBCut=as.f
actor(case_when(YearBuilt>=1800 & YearBuilt<1850 ~ "1800-1850",
YearBuilt>=1850 & YearBuilt<1900 ~ "1850-1900",
YearBuilt>=1900 & YearBuilt<1950 ~ "1900-1950",
YearBuilt>=1950 & YearBuilt<2000 ~ "1950-2000",YearBuilt>=2000  ~ "2000+"))) %>%
  ggplot(aes(x=fct_reorder(YBCut,GarageArea) ,y=GarageArea,fill=factor(YBCut)))+
  geom_boxplot()+
  coord_flip()+
  facet_wrap(~HouseStyle)+
  labs(x ="Year Built", y = "Garage Area", fill="Year Built")
```



Since 2000 on average the garage area has been increased. Before that we can say there was a time interval 1900-1950 in which garage area decreased on average compared to previous time intervel, 1850-1900. But after this reduction, we observe that garage area was increaded on average during time interval 1950-2000.