hw2_estes

Andrew Estes

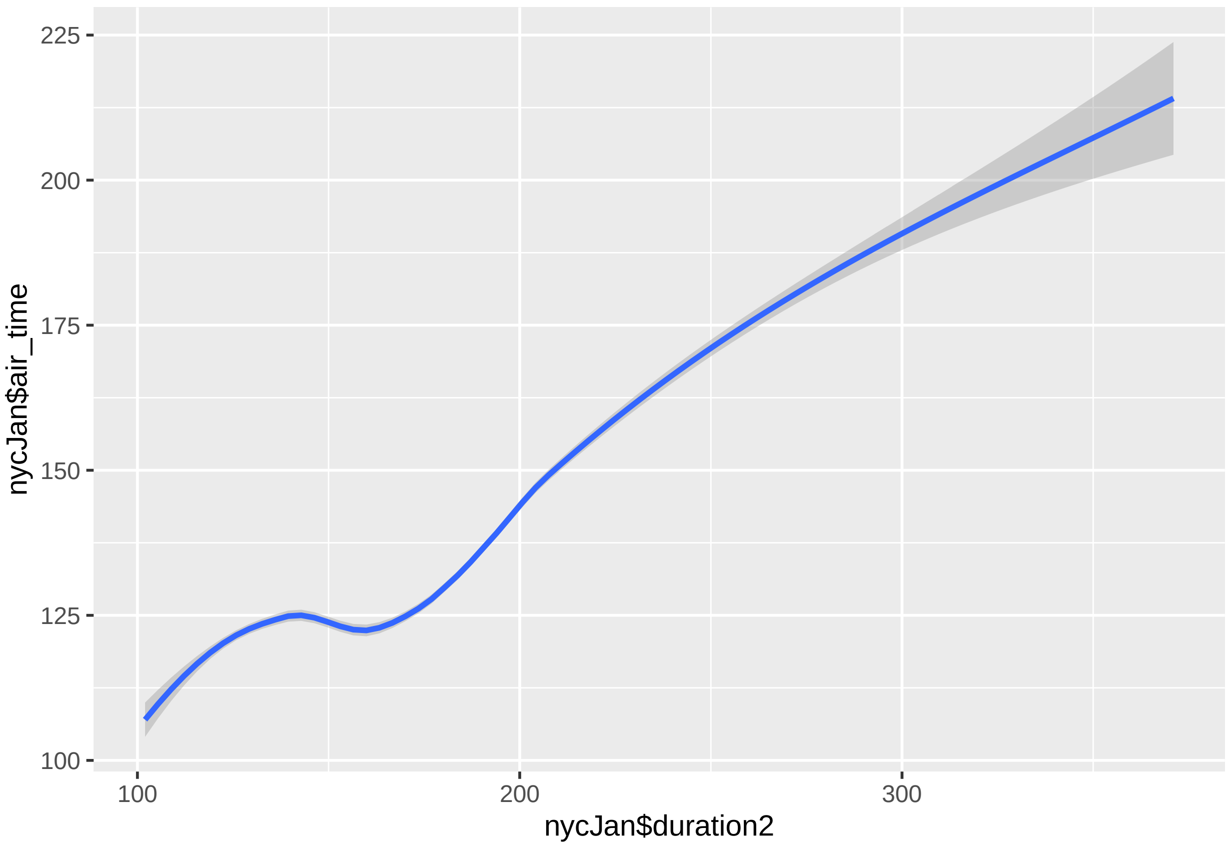
11/6/2021

```
library(nycflights13)
library(tidyverse)
library(lubridate)
library(AmesHousing)
```

```
#this creates the nycJan variable.
#It calls the flights dataframe for futher arguments
nycJan <- flights %>%
#the first argument is the month. We filter the flights df to only January
  filter(month==1) %>%
#this second argument changes the format of time for departures (517 vs 5:17)
  mutate(dep_time = dep_time/100) %>%
    unite("dep", c(year, month, day, dep_time), sep="/", remove = FALSE) %>%
    mutate(dep = ymd_hm(dep, tz = "America/New_York", quiet= TRUE)) %>%
    filter(!is.na(dep)) %>%
#this third argument changes the format of time for arrivals (740 vs 57:40)
  mutate(arr_time = arr_time/100) %>%
    unite("arr", c(year, month, day, arr_time), sep="/", remove = FALSE) %>%
    mutate(arr = ymd_hm(arr, tz = "America/New_York", quiet= TRUE)) %>%
    filter(!is.na(dep)) %>%
#the fourth argument only selects flights with a destination in the central timezone
  filter(dest == "ORD" | dest == "DSM" | dest == "STL" | dest =="MCI" | dest == "MDW") %>%
#this fifth argument calculates the flight duration in minutes
#air time is also in minutes
  mutate(duration = (arr_time - dep_time)*100) %>%
#the duration variable had negative numbers due to timing issues arrival time
#being 0130 (as in 130 in the am) with departure of 2330 (1130 pm)
  mutate(duration2 = ifelse(test = (duration < 0),
                            yes = duration + 2400,
                            no = duration))
```

Below is the graph showing the close relationship between air time and duration that widens as duration/air_time increase

```
ggplot(aes(x=nycJan$duration2,y=nycJan$air_time),data=nycJan) +
  geom_smooth(method = "loess")
```



```r
nycCommon <- flights %>%
  count(dest) %>%
  filter(n > 500)

nycCommon2 <- nycCommon[order(-nycCommon$n), ]
nycCommon2
```

```
## # A tibble: 72 x 2
##    dest      n
##    <chr> <int>
##  1 ORD   17283
##  2 ATL   17215
##  3 LAX   16174
##  4 BOS   15508
##  5 MCO   14082
##  6 CLT   14064
##  7 SFO   13331
##  8 FLL   12055
##  9 MIA   11728
## 10 DCA    9705
## # ... with 62 more rows
```

```r
nycKirksville <- nycCommon2 %>%
  filter(dest == "ORD" | dest == "DSM" | dest == "STL" | dest =="MCI" | dest == "MDW")
nycKirksville
```

```
## # A tibble: 5 x 2
##   dest      n
##   <chr> <int>
## 1 ORD   17283
## 2 STL    4339
## 3 MDW    4113
## 4 MCI    2008
## 5 DSM     569
```

This is not surprising. O'Hare is one of the 5 busiest airports in the US. DSM is the least busy airport in the list and is the smallest airport of the five listed. The middle three are all small-to-mid sized airports and fit correspondingly in the chart above.

```r
Ames <- make_ordinal_ames()
AmesNeigh <-  fct_unique(Ames$Neighborhood)

b.data <- str_count(AmesNeigh, "[bB]")
#b.data
sum(b.data)
```

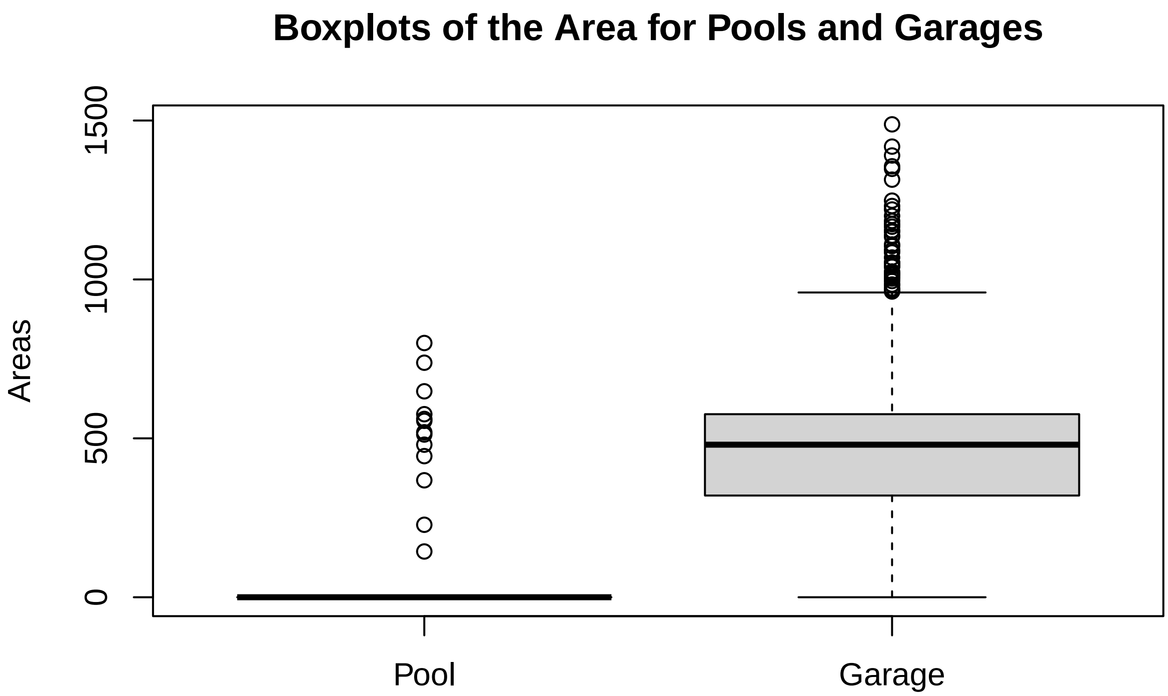
```
## [1] 7
```

```r
space.data <- str_count(AmesNeigh, fixed('_'))
#space.data
sum(space.data)
```

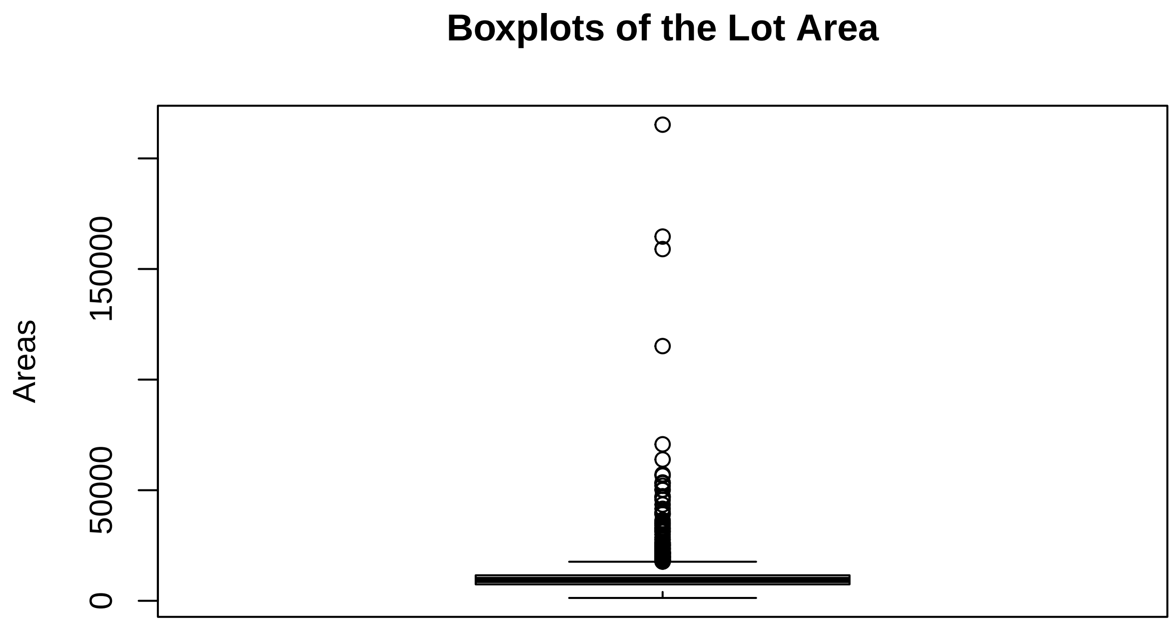
```
## [1] 23
```

```r
Ames2 <- Ames %>%
  mutate(Total_SF =
           Total_Bsmt_SF +
           Gr_Liv_Area +
           Garage_Area +
           Wood_Deck_SF +
           Open_Porch_SF +
           Enclosed_Porch +
           Three_season_porch +
           Screen_Porch
  )


Pool = Ames2$Pool_Area
Garage = Ames2$Garage_Area
Lot = Ames2$Lot_Area
bp <- boxplot(Pool, Garage,
        main = "Boxplots of the Area for Pools and Garages",
        ylab = "Areas",
        names = c("Pool", "Garage"))
```


Boxplots of the Area for Pools and Garages
Areas
0
500
1000
1500
Pool
Garage

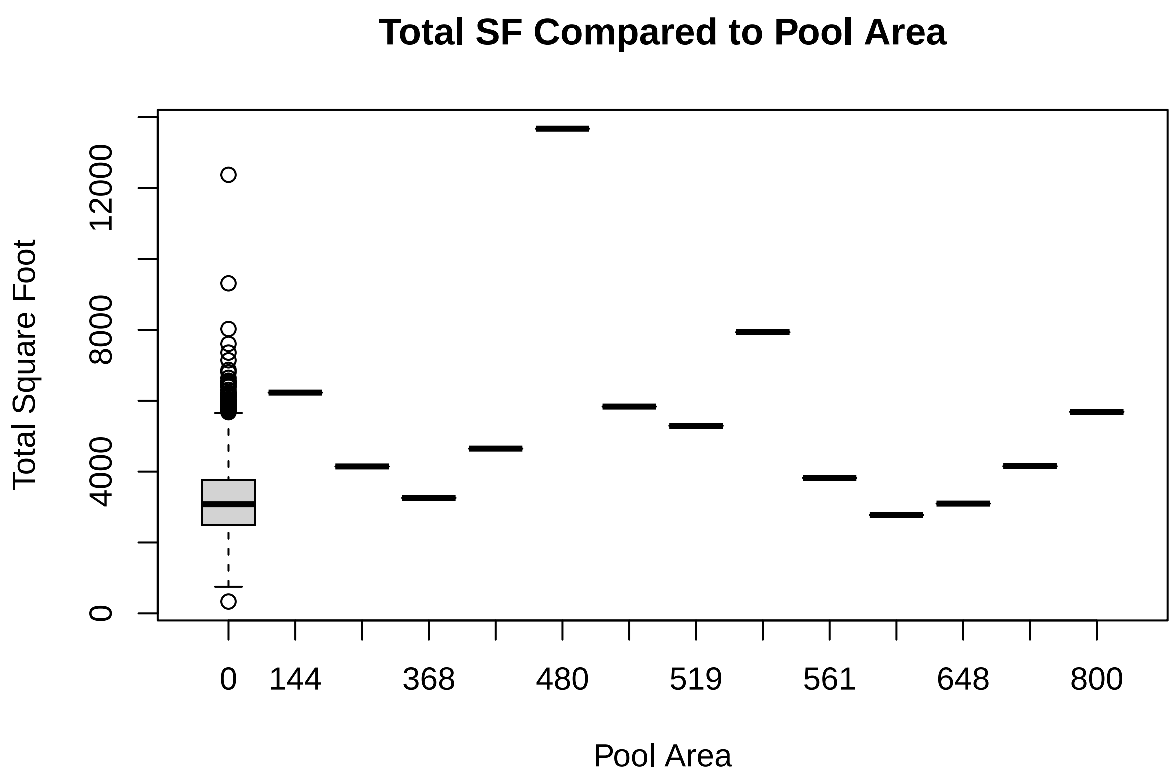
```r
bp.ots <- boxplot(Lot,
        main = "Boxplots of the Lot Area",
        ylab = "Areas",
        names = c("Lot"))
```



The pool and garage boxplots are to be expected. Most houses do not have pools so any pool house with a pool would likely be an outlier. As for the garage, 500 sqft is right around the average size of most 2-car garages. This data also makes sense. The outliers are more interesting than the pool outliers. I would hypothesize that the garage areas correlate to additional living/work space

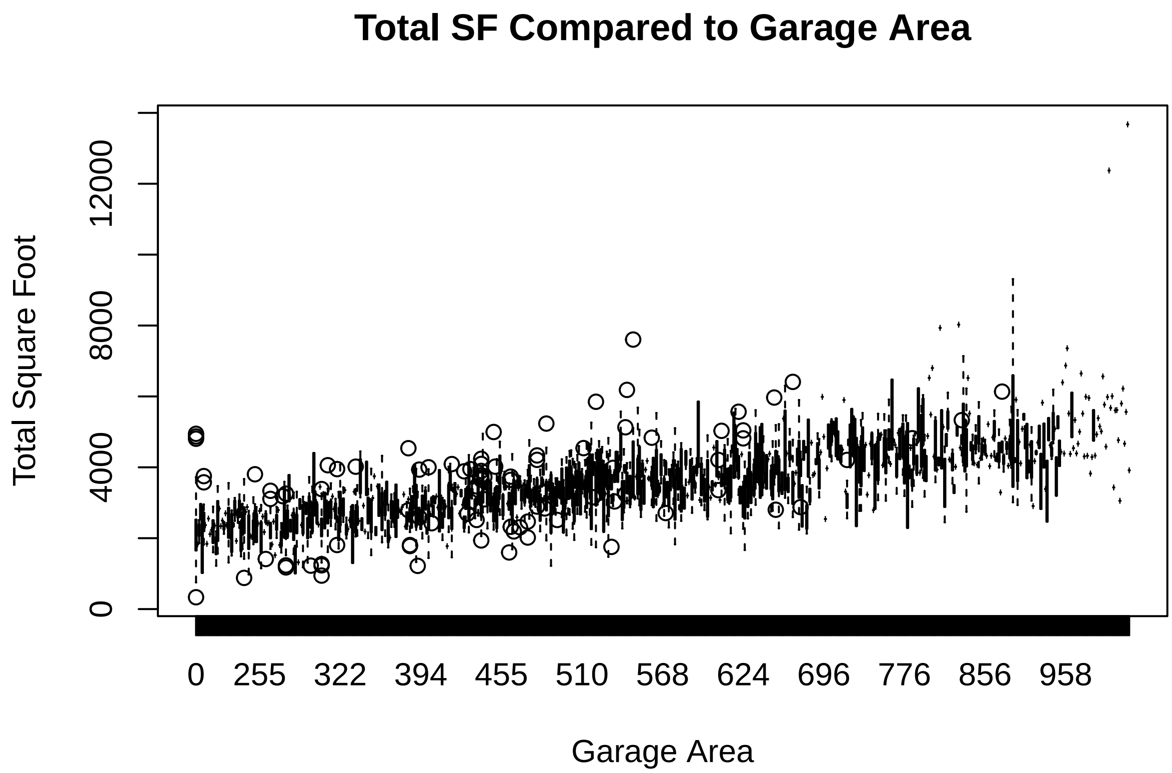
For my additional research I used the boxplot functions to create a neat correlation graph between the areas and total square feet of a property. The Total Sqare Feet excludes Lot Area as I wanted to focus on living/habitable space

```
boxplot(Total_SF~Pool_Area,data=Ames2, main="Total SF Compared to Pool Area",
        xlab="Pool Area", ylab="Total Square Foot")
```



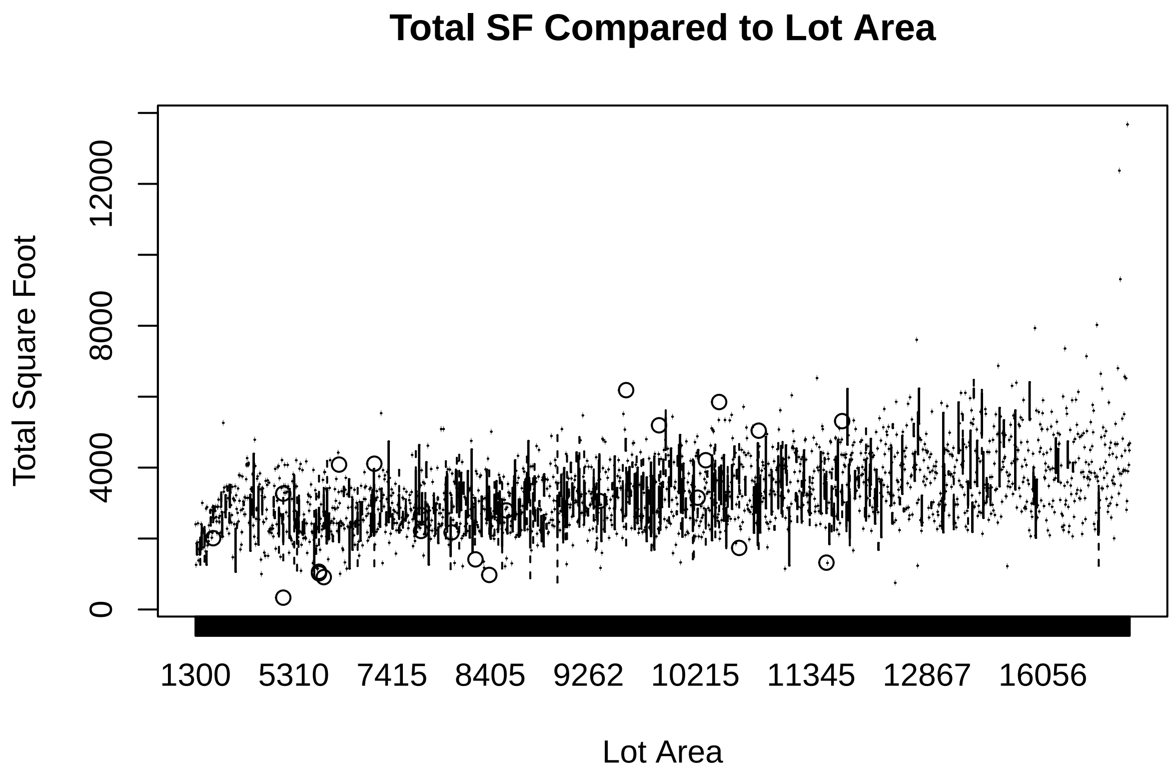
The pool area is obviously very scattered. Further analysis would require removing all data points with 0 square feet in Pool_Area

```
boxplot(Total_SF~Garage_Area,data=Ames2, main="Total SF Compared to Garage Area",
        xlab="Garage Area", ylab="Total Square Foot")
```



This is a more interesting graph. It shows a clear positive correlation between an increase in garage size and total square feet This is to be expected since garage area is a component of total square feet

```
my.bp <- boxplot(Total_SF~Lot_Area,data=Ames2, main="Total SF Compared to Lot Area",
        xlab="Lot Area", ylab="Total Square Foot")
```



This is a surprising graph. There are lots of land 5x as large as other houses but the total square feet of livable space remains roughly the same. There are some outliers of course, but predominantly the average remains the same regardless of lot area.