Square footage vs Sale price in Manhattan

Shravan Kuchkula, Ryan Khaleghi, Hieu Nguyen, Aaron Faltesek

2/28/2017

DATA GATHERING

The Department of Finance's Rolling Sales files lists properties that sold in the last twelve-month period in New York City for all tax classes. These files include the neighborhood, building type, square footage and other data.

Data set can be found here: http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

Packages used: install the gdata and plyr packages and load in to R. library(plyr) library(gdata)

set the current working directory

```
setwd("/Users/Shravan/R/projects/DS_6306_Unit6_Pract2/Paper")
```

read in the manhattan rolling sales data set and store it in a data frame.

```
mh <- read.xls("../Analysis/Data/rollingsales_manhattan.xls",skip=4,header=TRUE)
```

DATA CLEANING:

1. Create a new variable SALES.PRICE.N which does not contain dollar sign and comma's in it. Convert it from factor to numeric.

```
mh$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", mh$SALE.PRICE))
```

2. Make all variable names to lower case

```
names(mh) <- tolower(names(mh))
```

3. Convert gross.square.feet, land.square.feet and year.built from factor to numeric values. Store them in new variables.

```
mh$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", mh$gross.square.feet))
mh$land.sqft <- as.numeric(gsub("[^[:digit:]]","", mh$land.square.feet))
mh$year.built <- as.numeric(as.character(mh$year.built))
```

4. Take a backup of these changes. Since we are going to make some changes to the data frame.

```
mh_backup <- mh
```

5. As we are interested in finding the relationships between square ft and sale price, it makes sense to clean up these variables. In this step, we will be removing all the observations that have gross.sqft = NA and sale.price.n = NA

```
# Remove all observations which don't have gross.sqft
mh <- mh[!is.na(mh$gross.sqft),]
# Remove all observations which don't have sale.price.n
mh <- mh[!is.na(mh$sale.price.n),]
```

6. Remove observations for which sale.price.n = 0

```
mh <- mh[(mh$sale.price.n != 0),]
```

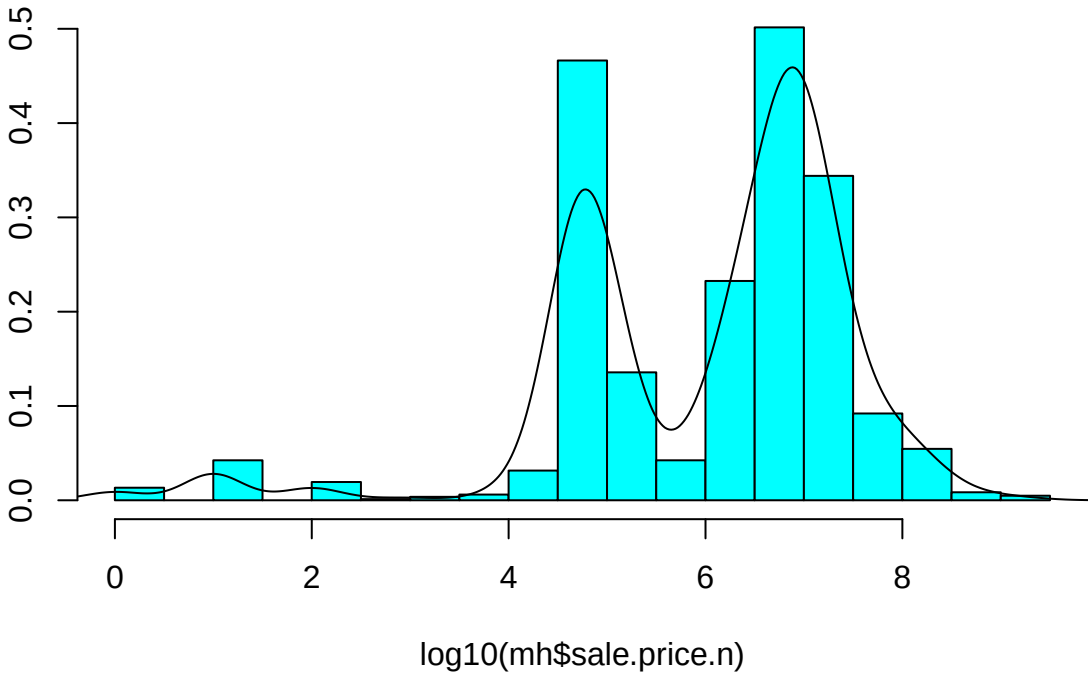
7. Remove observations for which gross.sqft = 0

```
mh <- mh[(mh$gross.sqft != 0),]
```

EXPLARATORY DATA ANALYSIS

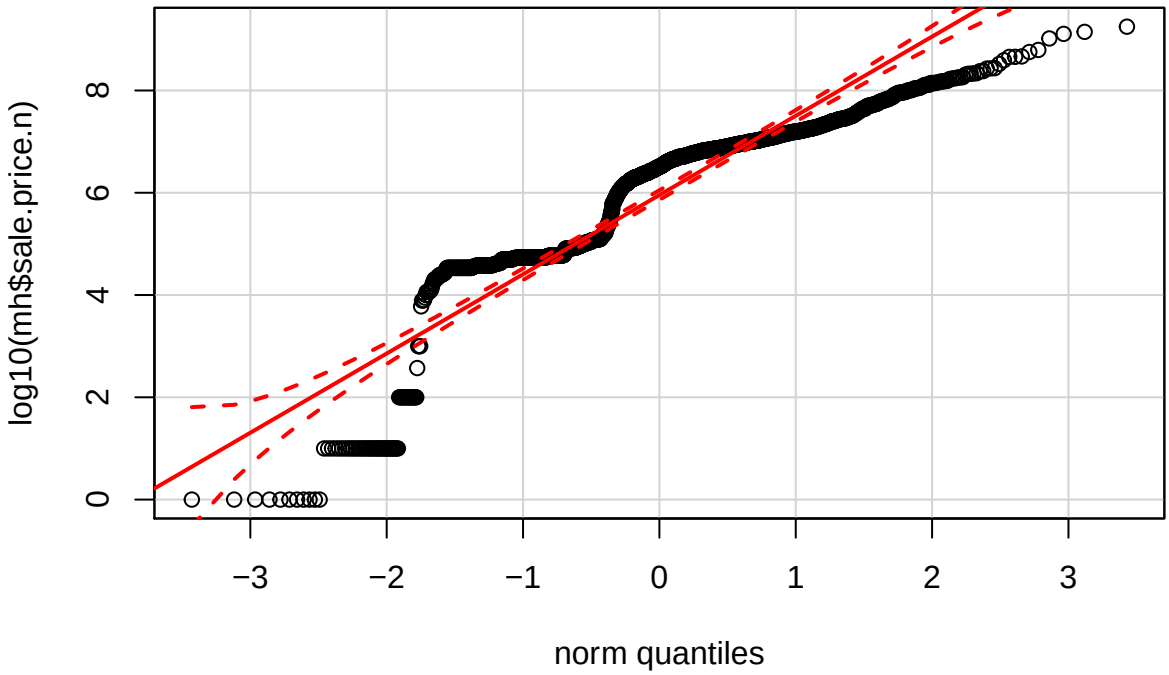
1. How is the sale.price.n data distributed for all neighbhorhoods ?

```
library(MASS)
truehist(log10(mh$sale.price.n))
lines(density(log10(mh$sale.price.n)))
```



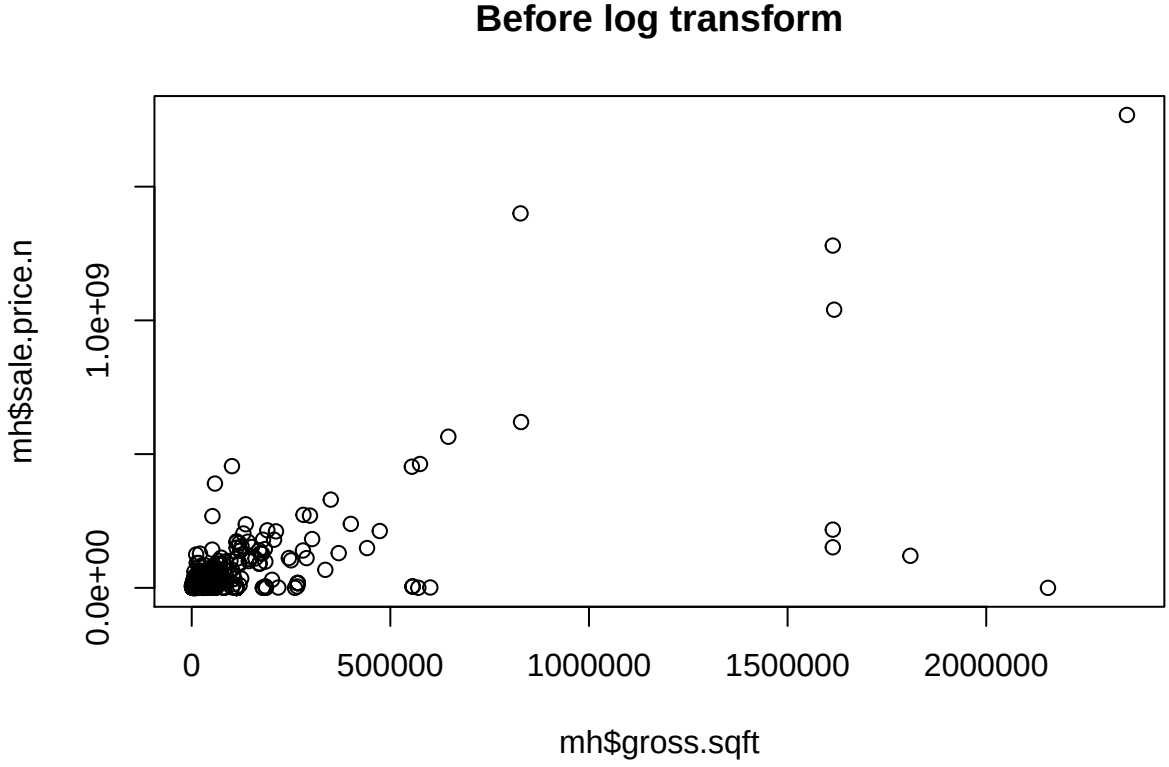
The sale price appears to be a bi-modal non-guassian distribution. A QQ-Plot also reveals that the sale price data is non-normal

```
library(car)
qqPlot(log10(mh$sale.price.n))
```

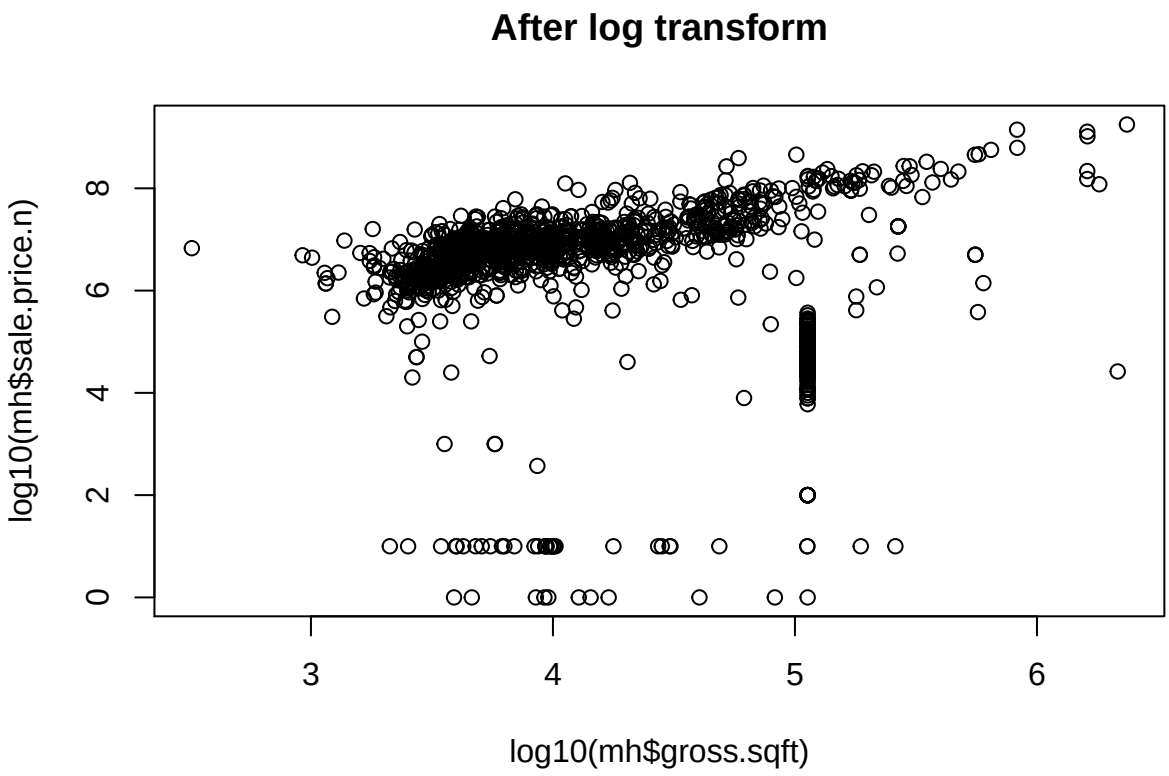


2. We then graphed a scatterplot of the sales price vs square feet to see if there was a trend. Again the data is heavily clustered around 0, but for those not in that blob, a slight linear trend can be observed.

```
op <- par(mfrow = c(1,1))
plot(mh$gross.sqft,mh$sale.price.n)
title("Before log transform")
```




```
plot(log10(mh$gross.sqft),log10(mh$sale.price.n))
title("After log transform")
```

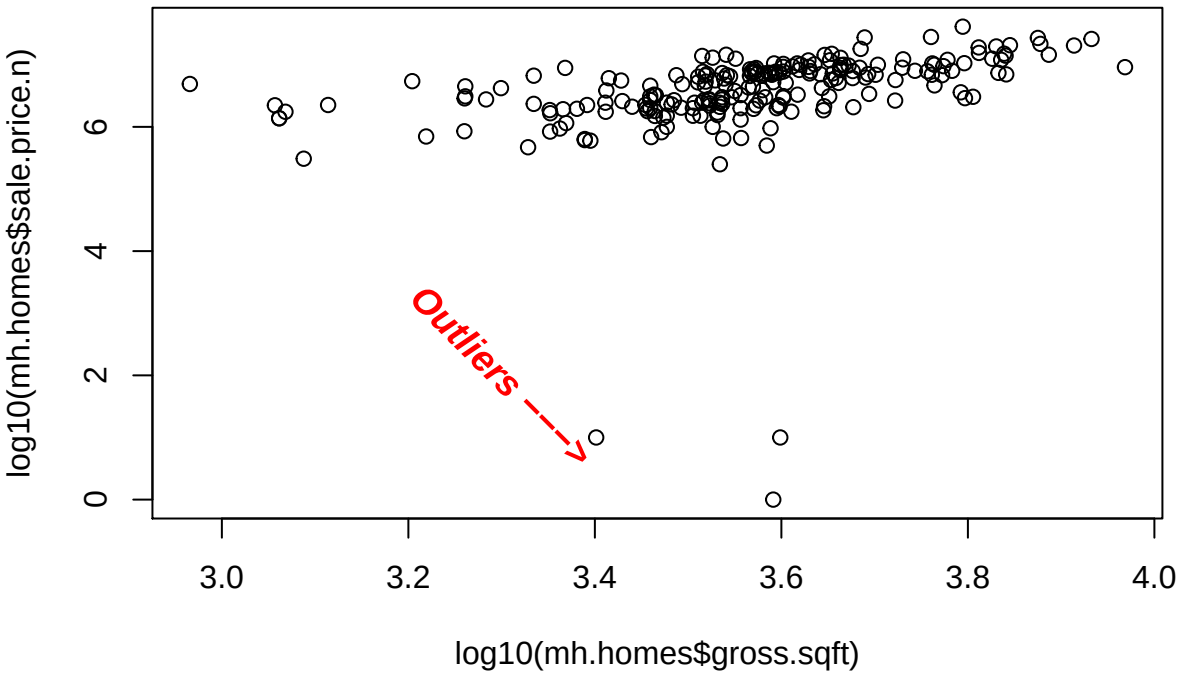



```
rm(op)
```

To investigate this trend further, we did a log10 transformation on the data set and examined this scatterplot. As we can see, the large grouping at the top section shows a line of data, but with the log transform and the scale, it's difficult to determine any trend. There is also a lot of data around 0 and above in a line that requires a large range on the axes to display completely and is not intuitive. A significant increase in the square footage of a dwelling should coincide with some price increase. For it to be flat is unusual. As well there is an outlier section at log10(x)=5 square feet where the price goes up but the square footage is identical. Some of these may not be sales or are simply missing data.

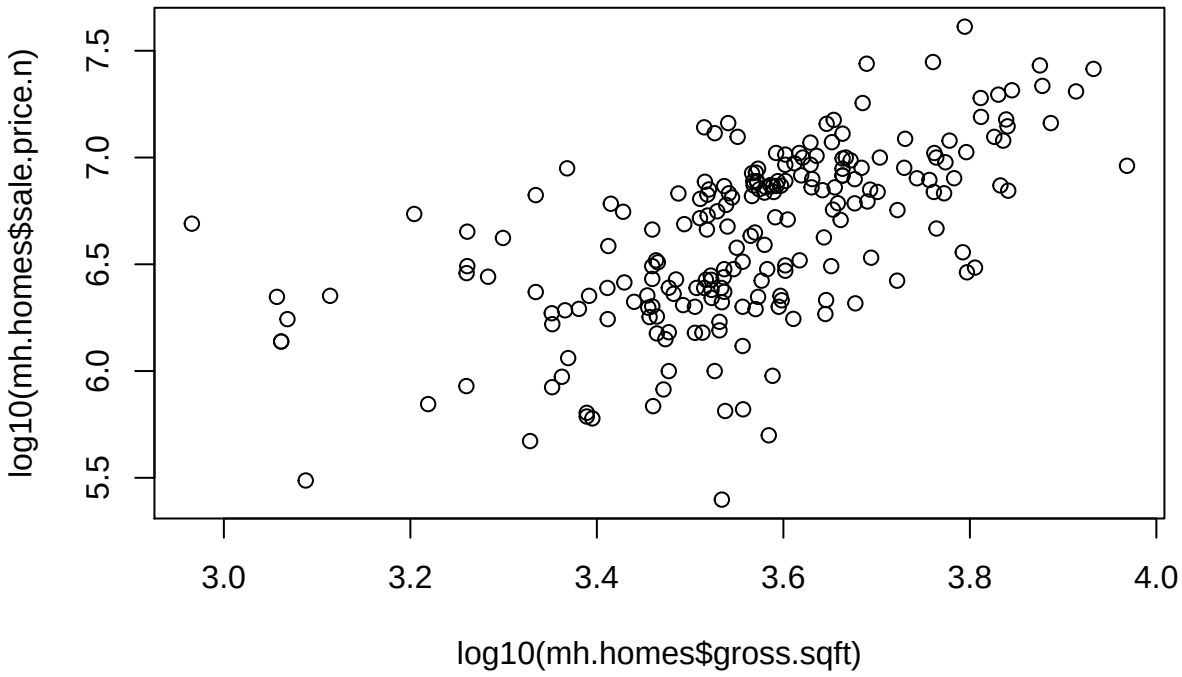
3. Since the data contains many different housing types and dwellings with a lot of missing data, we decided to focus on 1, 2, and 3 family homes and filter the data to exclude the others. This produces a much cleaner grouping around log10(x)=6 sales price, with some outliers.

```
mh.homes <- mh[which(grepl("FAMILY",mh$building.class.category)),]
plot(log10(mh.homes$gross.sqft),log10(mh.homes$sale.price.n))
text(x = 3.3, y = 2, labels = "Outliers --->", col = "red", cex = 1.2, font=4, srt = -45)
```



Notice the outliers in this scatterplot. Filtering these outliers out, since having a sale priceless than 5 is extremely unlikely, produces a graph of what are real sales of 1, 2, and 3 family homes on a log10 scale. We can see a distinct generally linear trend that the log of sales price increases with the log of square footage. We can also see that there is a significant amount of price variability for each level of square footage. As square footage increases, however, the variability decreases. This may simply be due to few homes at house sizes that are very large, and with real estate valuations tied to comparable home prices, this may lead to this clustering. There are definite pockets of clustering in the data, and further analysis is needed to determine if these are the same area and type, as above, that causes similar pricing, or it is due to another factor.

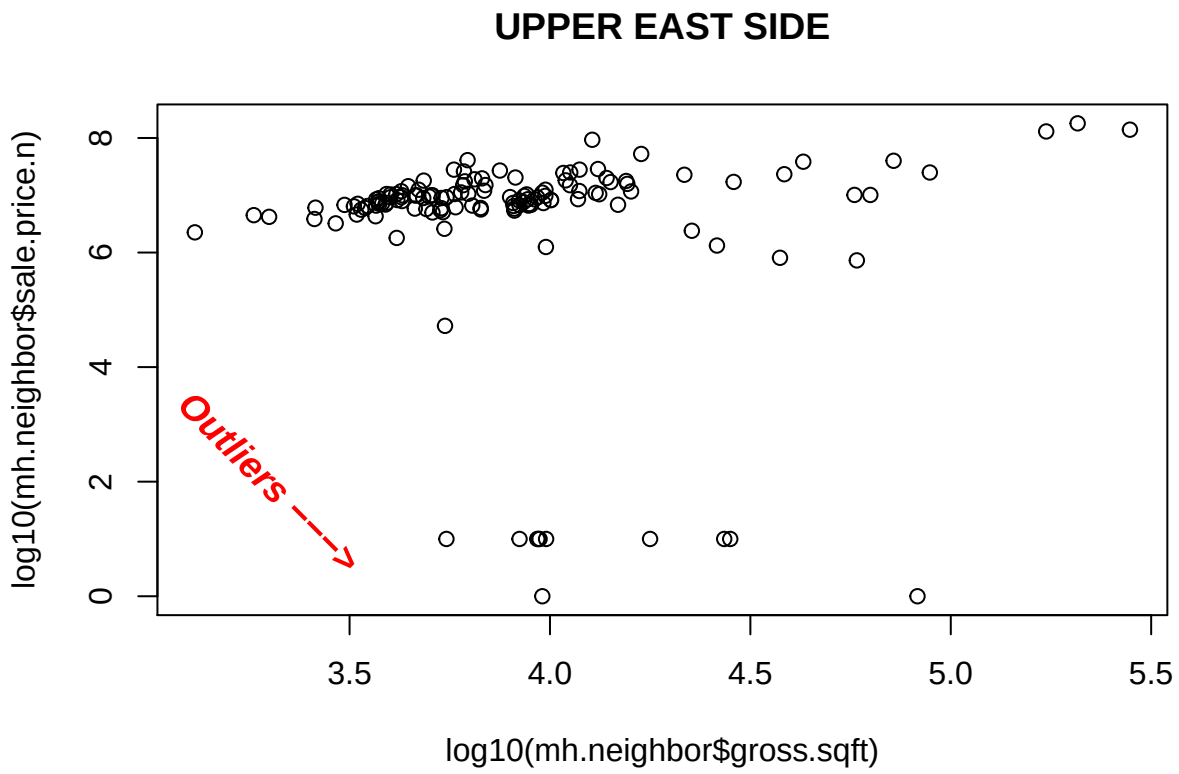
```
## remove outliers that seem like they weren't actual sales
mh.homes$outliers <- (log10(mh.homes$sale.price.n) <=5) + 0
mh.homes <- mh.homes[which(mh.homes$outliers==0),]
plot(log10(mh.homes$gross.sqft),log10(mh.homes$sale.price.n))
```



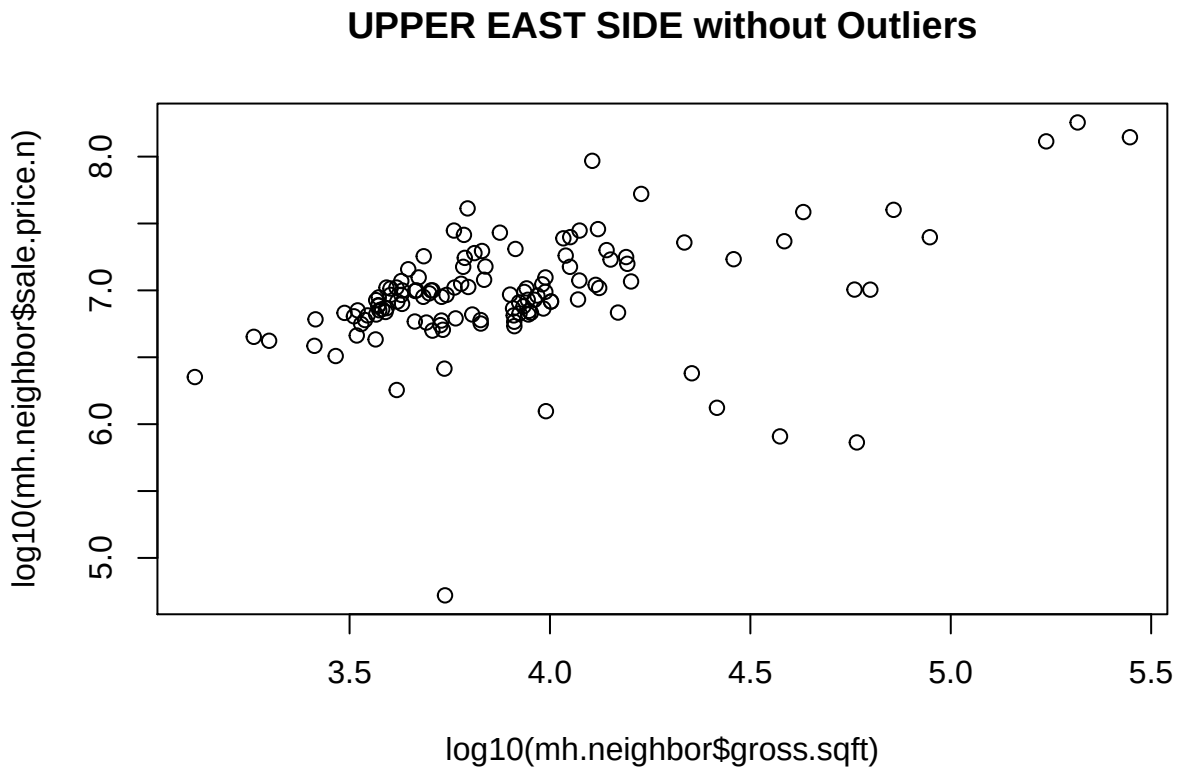
Interpretation: The data suggests that there is a positive linear correlation between gross square feet and sale price for 1,2,3 family dwellings accross all neighbhorhoods in Manhattan. In conclusion, we can generally say that for 1,2, and 3 bedroom homes in the Manhattan area, that there is a generally linear relationship between home sale price and square footage, with a slope on the log10 scale of 2.5.

4. We are intereseted in knowing the trend for UPPER EAST SIDE neighbhorhood.

```
# Upper east side neighbhorhood
op <- par(mfrow = c(1,1))
mh.neighbor <- mh[which(grepl("UPPER EAST SIDE", mh$neighborhood)),]
plot(log10(mh.neighbor$gross.sqft), log10(mh.neighbor$sale.price.n))
title("UPPER EAST SIDE")
text(x = 3.3, y = 2, labels = "Outliers --->", col = "red", cex = 1.2, font=4, srt = -45)
```




```
mh.neighbor$outliers <- (log10(mh.neighbor$sale.price.n) <=4) + 0
mh.neighbor <- mh.neighbor[which(mh.neighbor$outliers==0),]
plot(log10(mh.neighbor$gross.sqft), log10(mh.neighbor$sale.price.n))
title("UPPER EAST SIDE without Outliers")
```



```
rm(op)
```

Interpretation: The data suggests that there is a positive linear correlation between gross square feet and sale price for UPPER EAST SIDE neighbhorhood.