rollingSalesManhattan

Shravan Kuchkula

2/28/2017

install the gdata and plyr packages and load in to R.

```
library(plyr)
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
```

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)
```

Cleaning the data:

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))
count(is.na(mh$SALE.PRICE.N))
```

```
##       x  freq
## 1 FALSE 18519
## 2  TRUE  1684
```

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),]
```

Exploratory Data Analysis

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

```
hist(log10(mh$sale.price.n), breaks = 100, col="green", border="blue")
```