Kevin Pan

Data Analytics (Level 6000)

Assignment4

Due: March. 25, 2021

1. **For any one of the Brooklyn, Manhattan, Queens sales datasets, perform the following:**

   a). Describe the type of patterns or trends you might look for and how you plan to model them. Describe any exploratory data analysis you performed. Include plots and other descriptions. Min. 2-3 sentences (2%)

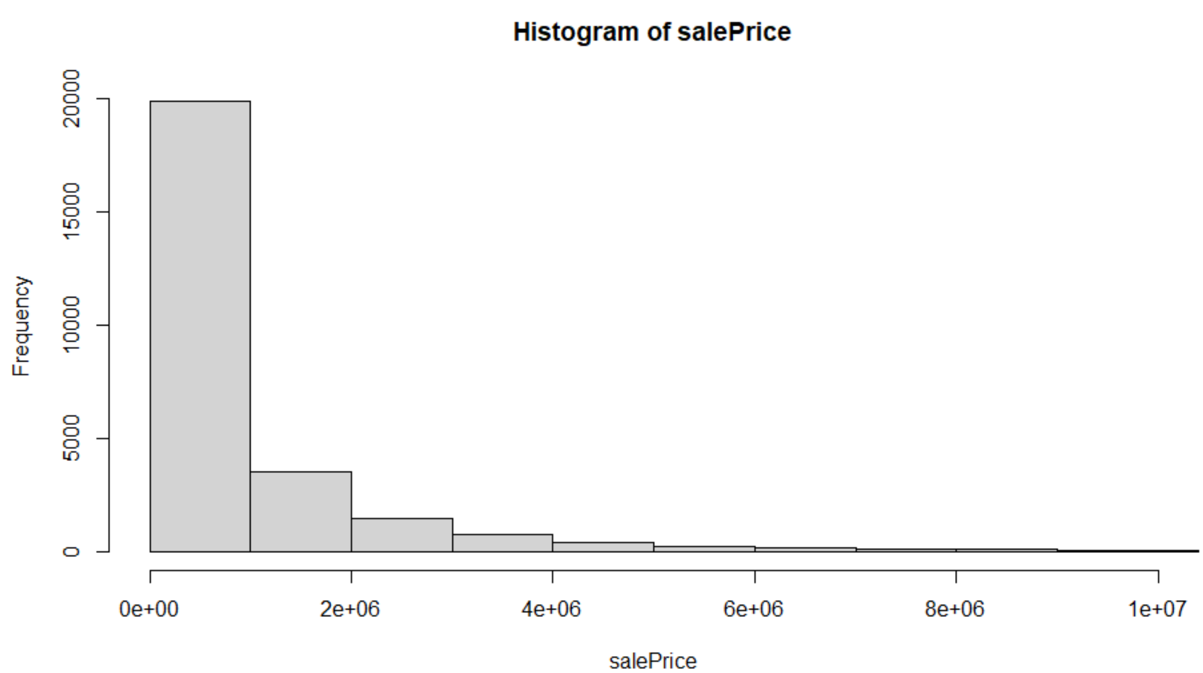
   The rollingsales_manhattan dataset is chosen for this assignment. Based on the dataset, my first hypothesis is that there could be a strong correlation between the sales price and the square feet. The store with a large square feet tends to sell items at a lower price. The second hypothesis is that some areas (zip code) of the Manhattan area could have a much higher sale price than others. For exploratory data analysis, I have gathered all the columns of the datasets and ran a brief summary on the entire dataset. Then I ran a summary and plotted histograms and boxplots of the attributes of interest.

```
> # retrieve the dataset
> library("readxl")
> getwd()
[1] "C:/Users/panke/Desktop/Graduate/DataAnalytics/Assign/HW4"
> mann <- read_excel("rollingsales_manhattan.xls", skip = 4)
> attach(mann)
> View(mann)
> summary(mann)
   BOROUGH  NEIGHBORHOOD      BUILDING CLASS CATEGORY TAX CLASS AT PRESENT
BLOCK         LOT
 Min.   :1   Length:27395       Length:27395          Length:27395        Min.   :   7   Min.   :   1.0
 1st Qu.:1   Class :character   Class :character      Class :character    1st Qu.: 877   1st Qu.:  37.0
 Median :1   Mode  :character   Mode  :character      Mode  :character    Median :1047   Median :1007.0
 Mean   :1                                                                Mean   :1110   Mean   : 741.8
 3rd Qu.:1                                                                3rd Qu.:1411   3rd Qu.:1233.0
 Max.   :1                                                                Max.   :2250   Max.   :9117.0
 EASE-MENT     BUILDING CLASS AT PRESENT   ADDRESS          APART\nMENT\nNUMBER    ZIP
CODE
 Mode:logical   Length:27395            Length:27395       Length:27395        Min.   :    0
 NA's:27395     Class :character        Class :character   Class :character    1st Qu.:10016
                Mode  :character        Mode  :character   Mode  :character    Median :10019
                                                                               Mean   :10029
                                                                               3rd Qu.:10027
                                                                               Max.   :10463
 RESIDENTIAL UNITS  COMMERCIAL UNITS   TOTAL UNITS      LAND SQUARE FEET   GROSS
SQUARE FEET   YEAR BUILT
 Min.   :  0.000   Min.   :  0.000   Min.   :  0.000   Min.   :      0.0   Min.   :       0   Min.   :   0
```

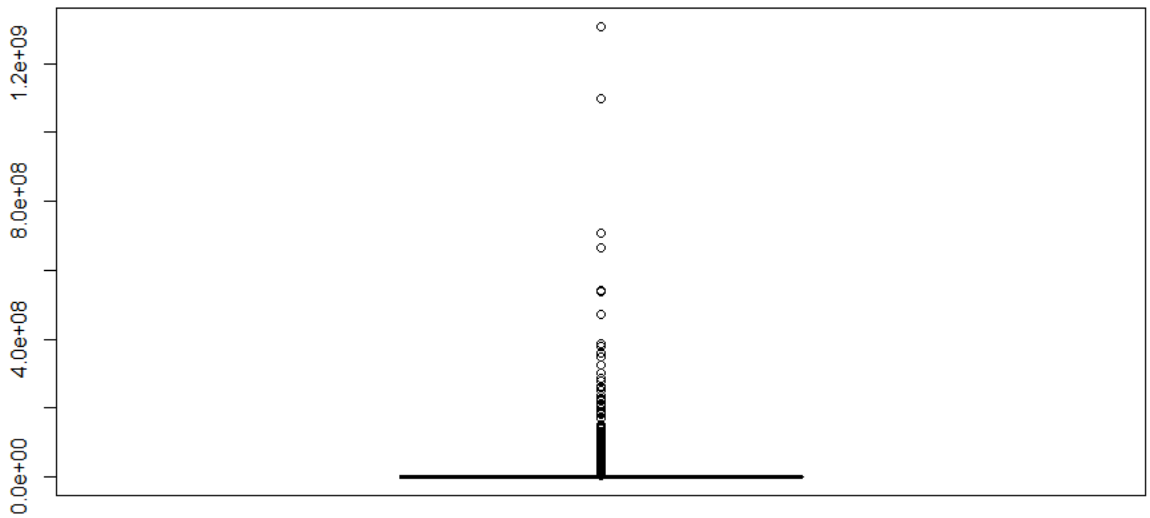
```
 1st Qu.:   0.000   1st Qu.:  0.000   1st Qu.:   0.000   1st Qu.:      0.0   1st Qu.:       0   1st Qu.:1900
 Median :   0.000   Median :  0.000   Median :   1.000   Median :      0.0   Median :       0   Median :1928
 Mean   :   1.766   Mean   :  0.375   Mean   :   2.289   Mean   :    965.7   Mean   :    9572   Mean   :1494
 3rd Qu.:   1.000   3rd Qu.:  0.000   3rd Qu.:   1.000   3rd Qu.:      0.0   3rd Qu.:       0   3rd Qu.:1973
 Max.   :1328.000   Max.   :604.000   Max.   :1349.000   Max.   :213650.0   Max.   :1970736   Max.   :2013
 TAX CLASS AT TIME OF SALE BUILDING CLASS AT TIME  OF SALE   SALE\nPRICE         SALE DATE
 Min.   :1.000             Length:27395                   Min.   :0.000e+00   Min.   :2012-08-01 00:00:00
 1st Qu.:2.000             Class :character               1st Qu.:0.000e+00   1st Qu.:2012-11-13 00:00:00
 Median :2.000             Mode  :character               Median :4.500e+05   Median :2013-01-17 00:00:00
 Mean   :2.488                                            Mean   :1.848e+06   Mean   :2013-01-31 14:59:03
 3rd Qu.:4.000                                            3rd Qu.:1.150e+06   3rd Qu.:2013-05-07 00:00:00
 Max.   :4.000                                            Max.   :1.308e+09   Max.   :2013-08-23 00:00:00
> colnames(mann)
 [1] "BOROUGH"                        "NEIGHBORHOOD"                   "BUILDING CLASS CATEGORY"
 [4] "TAX CLASS AT PRESENT"           "BLOCK"                          "LOT"
 [7] "EASE-MENT"                      "BUILDING CLASS AT PRESENT"      "ADDRESS"
[10] "APART\nMENT\nNUMBER"            "ZIP CODE"                       "RESIDENTIAL UNITS"
[13] "COMMERCIAL UNITS"               "TOTAL UNITS"                    "LAND SQUARE FEET"
[16] "GROSS SQUARE FEET"              "YEAR BUILT"                     "TAX CLASS AT TIME OF SALE"
[19] "BUILDING CLASS AT TIME OF SALE" "SALE\nPRICE"                    "SALE DATE"
> # initializing variables
> neighorhood <- mann$NEIGHBORHOOD
> zipCode <- mann$`ZIP CODE`
> salePrice <- mann$`SALE\nPRICE`
> saleDate <- mann$`SALE DATE`
> sqrFeet <- mann$`GROSS SQUARE FEET`

> # hypothesis: examine which neighborhood has the highest salePrice
> summary(salePrice)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.000e+00 0.000e+00 4.500e+05 1.848e+06 1.150e+06 1.308e+09
> hist(salePrice)
```

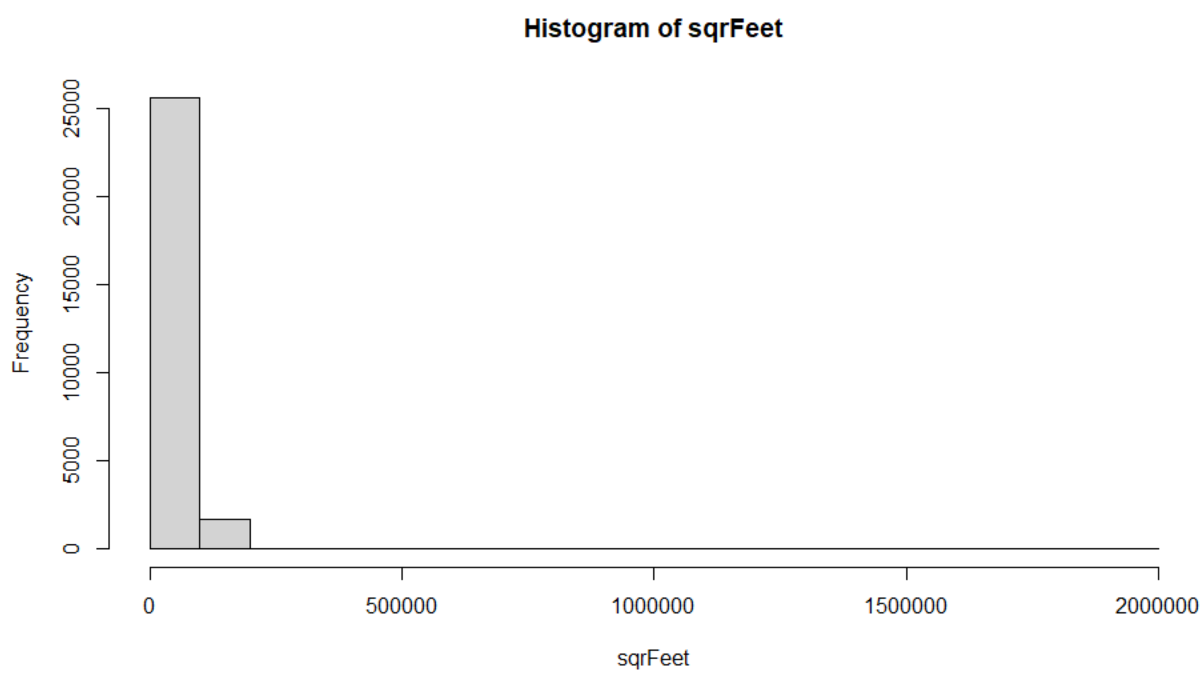
> hist(salePrice, xlim=c(0,1.0e+7), breaks=1000)



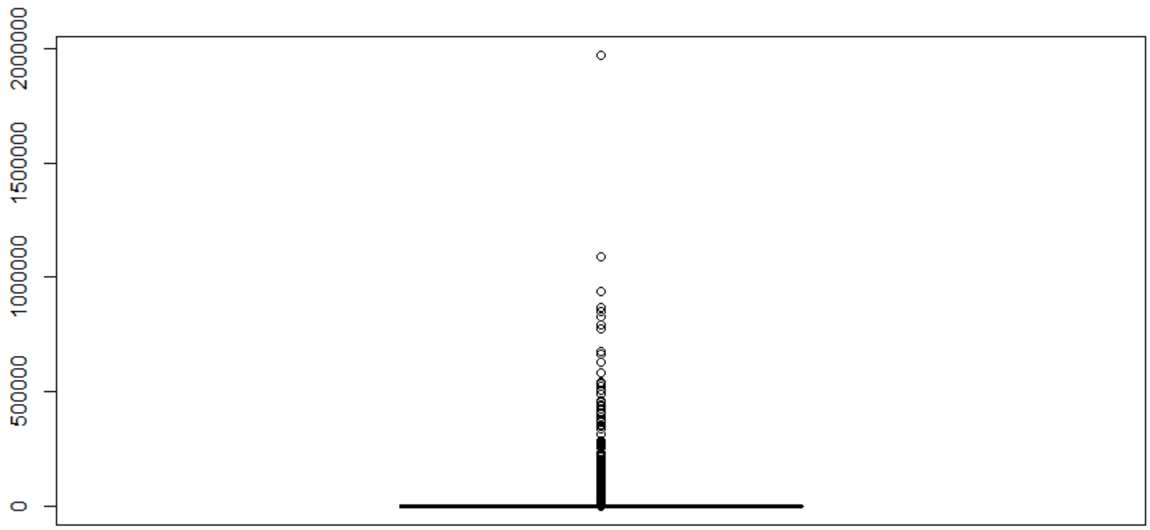
> boxplot(salePrice)




```
> summary(sqrFeet)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0       0    9572       0 1970736
> hist(sqrFeet)
```




```
> boxplot(sqrFeet)
```



b). Pick one or more models (these need not be restricted to the models you've learned so far [multivariate regression, KNN, K-Means]) to explore the chosen data. Interpret the model fits and indicates significance. Describe any cleaning you had to do and why. Min. 2-3 sentences (3%)

The data frame is built upon (salePrice, neighborhood, zipCode, saleDate, sqrFeet) and for data cleaning, rows with na values or 0 are removed. Those values are removed because the sale price and square feet is not possible to be 0. It's likely due to failure of gathering data. Then the data is split into testing and training sets.

For the model, multivariate regression is used for analysis. The goal is to examine which variable in the data frame has the most significant effect on the salePrice.

```
# data cleaning
library(dplyr)
df <- data.frame(salePrice, neighorhood, zipCode, saleDate, sqrFeet)
df <- na.omit(df)
df <- df[df$salePrice != 0,]
df = filter(df,data$salePrice != 0,)
df <- df[df$sqrFeet != 0,]
df = filter(df,data$sqrFeet != 0,)
View(df)
```

	salePrice	neighorhood	zipCode	saleDate	sqrFeet
19	3150000	ALPHABET CITY	10009	2013-03-06	3084
22	3650000	ALPHABET CITY	10009	2012-09-06	9345
23	895250	ALPHABET CITY	10009	2012-10-25	13002
25	283	ALPHABET CITY	10009	2013-04-18	5852
26	3500000	ALPHABET CITY	10009	2012-10-16	9071
27	13185684	ALPHABET CITY	10009	2013-01-31	18212
31	3810602	ALPHABET CITY	10009	2012-10-26	17664
32	7333333	ALPHABET CITY	10009	2013-04-09	6975
33	7333333	ALPHABET CITY	10009	2013-04-09	6875
34	7333333	ALPHABET CITY	10009	2013-04-09	7110
35	7000000	ALPHABET CITY	10009	2013-04-09	8975
37	12603963	ALPHABET CITY	10009	2013-01-31	15162
38	8500000	ALPHABET CITY	10009	2013-03-14	11000
40	8892981	ALPHABET CITY	10009	2013-01-31	15428
41	9528194	ALPHABET CITY	10009	2013-01-31	15428
42	4653771	ALPHABET CITY	10009	2013-01-31	10010
43	4653771	ALPHABET CITY	10009	2013-01-31	10010
46	9600000	ALPHABET CITY	10009	2013-05-07	9348
140	2800000	ALPHABET CITY	10009	2013-07-19	4140

```
library(caTools)
split = sample.split(df$salePrice, SplitRatio = 0.7)
training_set=subset(df,split==TRUE)
test_set=subset(df,split==FALSE)
```

2. For your chosen dataset:

a). Apply the model(s) to predict quantities of interest (that you choose). Describe (contingency table) or plot the predictions. Min. 2-3 sentences (6000-level 3%)

The results have a low p-value (2.2e-16). The residual standard error is also really high. Upon a closer examination on the attributes, zip codes seem to contribute to the prediction of the sale price the most, this is likely due to some stores with higher price clusters in the same area in manhattan. So their zip codes are the same or a few numbers off from each other.

```
# using the models to predict
regressor=lm(formula=training_set$salePrice~., data=training_set)
summary(regressor)
```

Call:

lm(formula = training_set$salePrice ~ ., data = training_set)

```
Residuals:
      Min        1Q    Median        3Q       Max
-231666686  -3410068  -1987144   1379892  592612057

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -3.718e+08  3.018e+08  -1.232   0.2180
neighorhoodCHELSEA                     6.434e+06  7.419e+06   0.867   0.3859
neighorhoodCHINATOWN                   1.141e+06  9.433e+06   0.121   0.9037
neighorhoodCIVIC CENTER               -2.013e+06  1.406e+07  -0.143   0.8862
neighorhoodCLINTON                     1.628e+06  8.448e+06   0.193   0.8472
neighorhoodEAST VILLAGE                2.172e+06  7.804e+06   0.278   0.7808
neighorhoodFASHION                    -3.485e+06  8.127e+06  -0.429   0.6681
neighorhoodFINANCIAL                  -4.492e+06  1.033e+07  -0.435   0.6636
neighorhoodFLATIRON                   -4.338e+07  8.512e+06  -5.097 3.76e-07 ***
neighorhoodGRAMERCY                    2.099e+07  8.798e+06   2.385   0.0172 *
neighorhoodGREENWICH VILLAGE-CENTRAL  3.938e+06  8.511e+06   0.463   0.6436
neighorhoodGREENWICH VILLAGE-WEST     6.104e+06  7.529e+06   0.811   0.4176
neighorhoodHARLEM-CENTRAL             -5.970e+06  6.630e+06  -0.901   0.3679
neighorhoodHARLEM-EAST                -4.425e+06  7.272e+06  -0.609   0.5429
neighorhoodHARLEM-UPPER               -7.345e+06  7.693e+06  -0.955   0.3398
neighorhoodHARLEM-WEST                -8.536e+06  8.910e+06  -0.958   0.3382
neighorhoodINWOOD                     -9.978e+06  9.507e+06  -1.050   0.2940
neighorhoodJAVITS CENTER               9.518e+06  1.406e+07   0.677   0.4985
neighorhoodKIPS BAY                    5.357e+05  1.176e+07   0.046   0.9637
neighorhoodLITTLE ITALY                2.089e+06  9.795e+06   0.213   0.8312
neighorhoodLOWER EAST SIDE             3.487e+06  8.231e+06   0.424   0.6719
neighorhoodMANHATTAN VALLEY           -6.278e+06  8.245e+06  -0.761   0.4465
neighorhoodMIDTOWN CBD                 1.465e+08  9.996e+06  14.653  < 2e-16 ***
neighorhoodMIDTOWN EAST                4.956e+06  8.439e+06   0.587   0.5571
neighorhoodMIDTOWN WEST               -4.844e+07  6.445e+06  -7.516 8.24e-14 ***
neighorhoodMORNINGSIDE HEIGHTS        -4.878e+06  2.085e+07  -0.234   0.8150
neighorhoodMURRAY HILL                 4.699e+06  8.018e+06   0.586   0.5579
neighorhoodSOHO                        1.867e+07  7.672e+06   2.433   0.0150 *
neighorhoodSOUTHBRIDGE                 2.844e+07  1.242e+07   2.289   0.0222 *
neighorhoodTRIBECA                     1.611e+07  1.002e+07   1.608   0.1080
neighorhoodUPPER EAST SIDE (59-79)     1.894e+06  7.079e+06   0.267   0.7891
neighorhoodUPPER EAST SIDE (79-96)    -5.408e+05  7.243e+06  -0.075   0.9405
neighorhoodUPPER WEST SIDE (59-79)    -8.968e+05  8.452e+06  -0.106   0.9155
neighorhoodUPPER WEST SIDE (79-96)    -1.776e+06  7.813e+06  -0.227   0.8202
neighorhoodUPPER WEST SIDE (96-116)   -7.582e+06  9.149e+06  -0.829   0.4074
neighorhoodWASHINGTON HEIGHTS LOWER  -9.110e+06  7.353e+06  -1.239   0.2155
```

```
neighorhoodWASHINGTON HEIGHTS UPPER  -9.811e+06  7.402e+06  -1.326   0.1851
zipCode                           2.289e+04  2.864e+04   0.799   0.4242
saleDate                          1.064e-01  6.949e-02   1.531   0.1260
sqrFeet                           4.330e+02  1.040e+01  41.638  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28110000 on 2170 degrees of freedom
Multiple R-squared:  0.6014,	Adjusted R-squared:  0.5942
F-statistic: 83.95 on 39 and 2170 DF,  p-value: < 2.2e-16
```

b). Examine the fit(s). Perform a significance test that is suitable for the variables you are investigating and describe the results. Min. 2-3 sentences (6000-level 3%)

One sample t-test is performed, the p-value is small (as expected from the results of the model). Based on the summary of the output, no variables in the model have high fits.

```
> t.test(df$salePrice)

	One Sample t-test

data:  df$salePrice
t = 11.219, df = 2764, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  7029890 10007769
sample estimates:
mean of x
  8518830
```

c). Discuss any observations you had about the datasets/ variables, other data in the dataset and/or your confidence in the result. Min 1-2 sentences (1%)

The confidence of the result is low. By using the multivariate regression model, no attributes in the dataframe seems to be directly proportional to the sale price. Perhaps some non-linear transformation will capture the features better.

3. 6000-level question (3%). Draw conclusions from this study – about the model type and suitability/ deficiencies. Describe what worked and why/ why not. Min. 4-5 sentences

The sale prices are not directly related to neighborhood, zipCode, saleDate, sqrFeet. This contradicts my hypothesis. My hypothesis states that the sale price has a high correlation with the square feet and the zipcode. Perhaps, the sale price is more closely related to some other attributes that are not included in the dataframe, or the sale price is just random. From the result, the deficiencies are high because the model showed a high significaticace, high error standard deviation, and low p-value.