Assignment 1 - Manhattan Project

Andresa de Andrade

Tuesday, February 17, 2015

Executive Summary -Problem Presentation;

-Data Cleaning;

-Exploratory Analysis;

-Statistics Methodology;

-Conclusion;

Problem Presentation

Due the high demand for properties in Manhattan, the city hall wants to anticipate/predict the prices for real state in order to prevent a new crises. In this case, it's necessary to understand what are the factors that contribute the price of the properties and how significant are those factors to build the final price.

In order to answer those question we will be working with the following variables:

1. Building Class
2. Zipcode
3. Number of residential units
4. Number of commercial units
5. Year that the building was built
6. Sale price
7. Sale date

We will also an unique variable called id to make it easier to describe.

Cleaning Data

First it's important to mention that we removed the first rows in the excel file and converted to a csv file to be more workable in R.

When we are cleaning the data, there are two options, exclude missing data and adapt/transform the missing data but keep the record in the database.

In this project we'll be applying both of them, but I'll start by the first one.

We will be deleting records that don't have a sale price and those that don't have a zipcode.

The reason is very simple, we can't predict a value that we don't have

So now we can work with our data and see if the analysis makes sense.

Exploratory Analysis

First let's see how the main statistics of this database looks like. The summary from R gives us the main descriptive stats.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 1.00e+00 3.68e+05 7.40e+05 2.56e+06 1.63e+06 1.31e+09
```

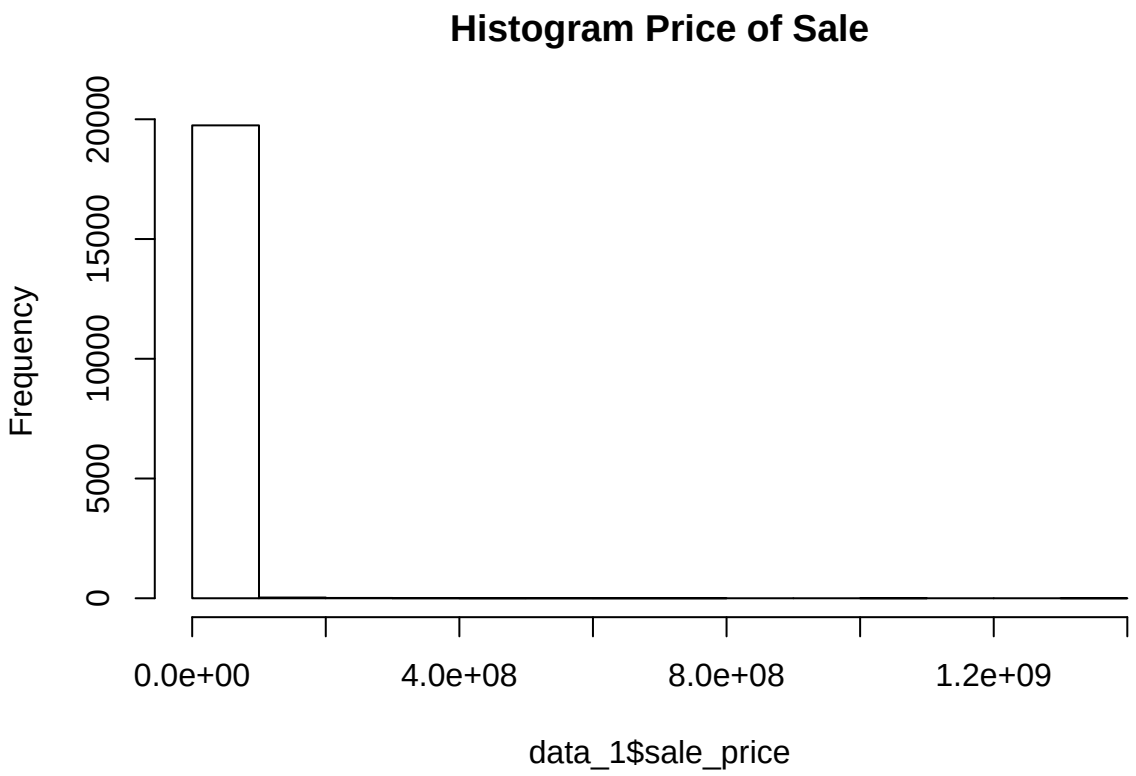
The minimum value that a property was sold is 1 dollar and the maximum is 1, and the max is 1.3B which suggests that we have a heavy presence of outliers in our dataset.

I'm not sure why we have the value 1 dollar as a price for a house (if it's truthful our a misspelling), but for now I'll consider as a true value.

To investigate this a little further let's plot the distribution for the price.

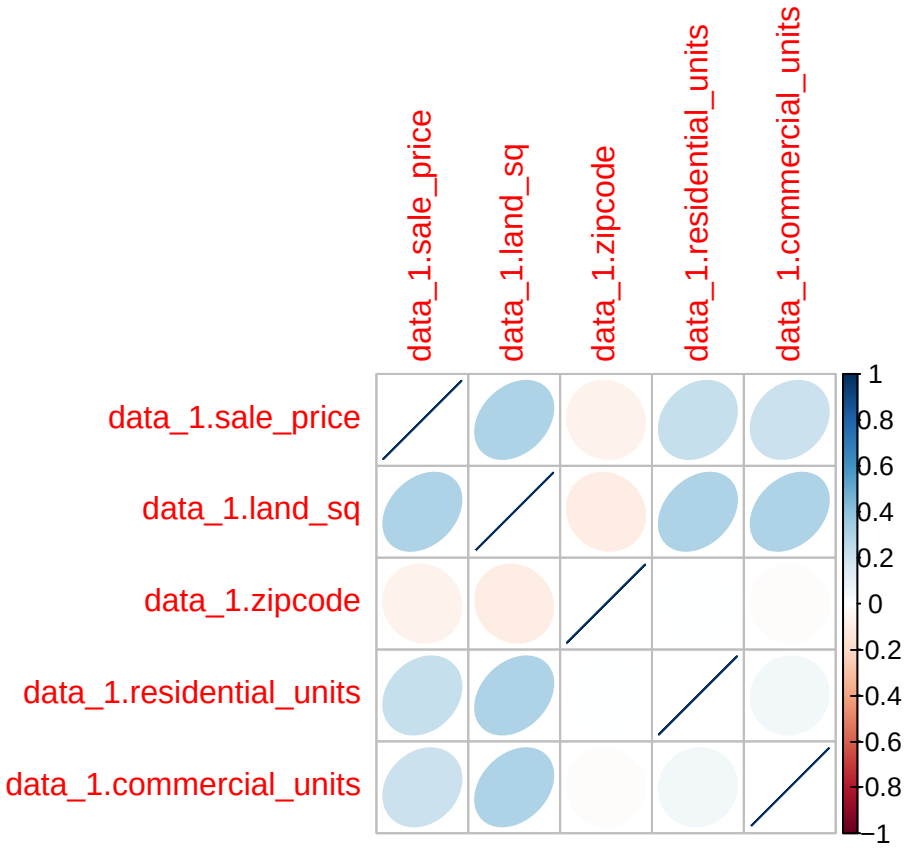
```
## Warning: NAs produced by integer overflow
```

As we can see, upper outliers are very far way from the rest of the data, this might cause some issues in the model because the linear model is not the best best approach when we are dealing with extreme values.



We notice that we have positive skewness in the histogram, so majority of our data is concentrated in the left part of our chart.

We should also check the correlation for our variables to understand how the complement each other




```
##                          data_1.sale_price data_1.land_sq data_1.zipcode
## data_1.sale_price                  1.00000        0.30587      -0.060974
## data_1.land_sq                     0.30587        1.00000      -0.095782
## data_1.zipcode                    -0.06097       -0.09578       1.000000
## data_1.residential_units           0.23250        0.30949       0.006523
## data_1.commercial_units            0.21321        0.30988      -0.013571
##                          data_1.residential_units data_1.commercial_units
## data_1.sale_price                        0.232502                 0.21321
## data_1.land_sq                           0.309489                 0.30988
## data_1.zipcode                           0.006523                -0.01357
## data_1.residential_units                 1.000000                 0.05168
## data_1.commercial_units                  0.051680                 1.00000
```

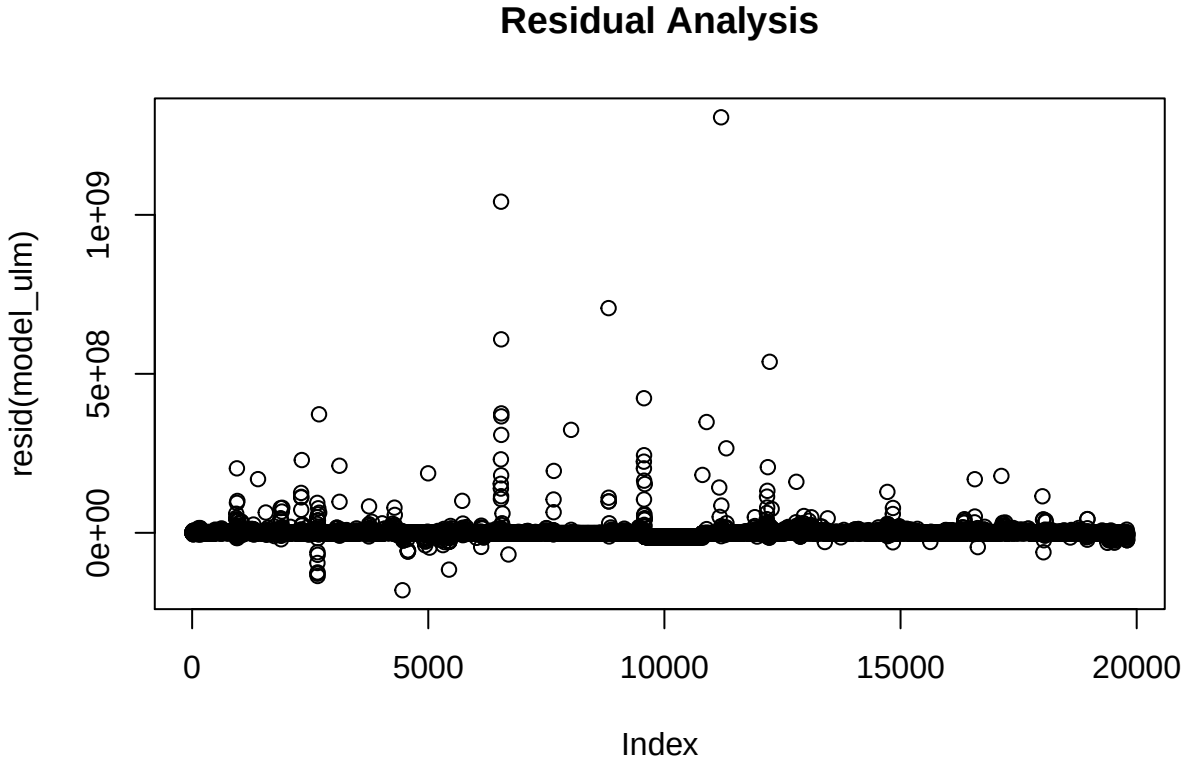
The price and the the size of the property in square feet appear to have a positive correlation, in other worlds the bigger the size, more expensive the value. The same conclusion may be applied for the number of commercial and residential units in a building, more units means higher the value.

Regression Model

Our first model is a univariate model using price versus land size in square feet:

```
##
## Call:
## lm(formula = sale_price ~ land_sq, data = data_1)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.81e+08 -8.46e+05 -4.76e+05  2.99e+05  1.31e+09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.15e+06   1.33e+05    8.64   <2e-16 ***
## land_sq     1.56e+03   3.45e+01   45.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18200000 on 19800 degrees of freedom
## Multiple R-squared:  0.0936, Adjusted R-squared:  0.0935
## F-statistic: 2.04e+03 on 1 and 19800 DF,  p-value: <2e-16
```



We have a model that has a p-value for the F test very small which suggests that model has a variance small enough to be accepted. For this model we have two coefficients, the intercept and the coefficient of land_square, and both are very small suggesting a significance to the model.

In this case we have a model that looks like price = 1.56x10^6 +1.53x10^3(land_size)

We need also check the residuals to ensure that are normaly distribuited. It's more less normal distribuited, because it has the presence of outliers.

Now we have our multivariate model using the price as dependent variable and land size, zipcode and residential and commercial units as indenpendent variables.

```
##
## Call:
## lm(formula = sale_price ~ land_sq + zipcode + residential_units +
```

```
##     commercial_units, data = data_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.60e+08 -7.06e+05 -3.19e+05  5.08e+05  1.31e+09
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.76e+07   1.61e+07    6.08  1.2e-09 ***
## land_sq            1.07e+03   3.75e+01   28.41  < 2e-16 ***
## zipcode           -9.65e+03   1.60e+03   -6.03  1.7e-09 ***
## residential_units  2.61e+05   1.13e+04   23.08  < 2e-16 ***
## commercial_units   1.19e+06   5.96e+04   20.03  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17800000 on 19797 degrees of freedom
## Multiple R-squared:  0.134,  Adjusted R-squared:  0.133
## F-statistic:  763 on 4 and 19797 DF,  p-value: <2e-16
```

Although the multivariate model is still good in terms of 1) Model adjustiment and 2)coefficients significance we need to highlight that the Rsquare is relatively low, in other worlds the model doesn't explain most of the variance of our data.

Cross Validation

With cross validation we are also able to test our data. So for this case we are partioning the datase into 10 pieces and using a part as the training and the other as testing.

```
## Warning: package 'caret' was built under R version 3.1.2
```

```
## Loading required package: ggplot2
```

summary(model_ulm2)

Conclusion

So from all of the content above we can say that land_square and neighborhood are very significant for the price of a property in Manhattan, the zipcode is related to the lower the zipcode more uptown is the neighborhood and therefore more valuable for the price.