Marketing Analytics Course 1.3

First I downloaded the data prep (dataprep1) Excel file from Edx. The dataprep file needs to be edited and the categorical Neighborhood column turned into numberic data. We do so by creating dummy variables with three columns for the different neighborhoods. We only need two of the columns to discern the infomation of the three neighborhood columns, so we remove one column. Now we save the Excel file as a column seperated values or csv file. Using a csv file makes it easy to upload our data into R.

Here we read our file into R and save it as dataframe named DataReal.df

```
RealData.df <- read.csv("edx_dataprep1.2.csv")
```

Now we check to see if our file has been downloaded correctly by printing our data frame.

```
RealData.df
```

```
##    Price1000 SizeSqFt LotAcre Year Bedrooms Bathrooms Neigborhood_Ladera
## 1       2488     2640    0.33 1962        3       3.0                   1
## 2       2688     1900    0.28 1959        3       2.0                   1
## 3       2750     1760    0.33 1956        3       2.0                   1
## 4       2698     2668    0.41 1988        4       3.0                   1
## 5       3998     3465    0.98 1963        4       3.0                   0
## 6      16500     6610    9.77 1957        5       5.0                   0
## 7       6895     4670    0.99 1988        4       5.0                   0
## 8      16988     6808   17.99 2001        5       8.0                   0
## 9       6595     5542    1.16 2015        5       6.0                   0
## 10     16000     3668   13.84 1955        5       4.0                   0
## 11      4695     3600    1.14 1964        5       3.0                   0
## 12      4900     3600    2.50 1953        3       2.0                   0
## 13      6800     4700    1.08 1959        3       4.0                   0
## 14      5695     3900    2.51 1989        4       4.0                   0
## 15      3000     1660    0.21 1955        3       2.0                   1
## 16      3305     2210    2.69 1954        3       2.0                   0
## 17      2880     1870    0.27 1959        4       2.0                   1
## 18      3800     3685    0.50 1986        4       3.0                   0
## 19      6300     5273    1.38 1957        6       7.0                   0
## 20      4461     2660    1.03 1959        4       4.0                   0
## 21     11250     8333    2.51 2008        4       4.0                   0
## 22      4845     4350    1.18 2011        5       3.0                   0
## 23      4300     2790    1.05 1958        5       3.0                   0
## 24      4650     2640    2.06 1954        3       3.0                   0
## 25      2800     2650    0.20 2004        3       4.0                   1
## 26      6598     4282    1.23 1951        5       6.0                   0
## 27      2750     2600    0.42 1959        3       3.0                   1
## 28      5235     2670    2.02 1957        3       4.0                   0
## 29      3700     2710    0.20 1952        4       3.0                   1
## 30      2745     2890    0.34 2007        4       4.0                   0
## 31      4025     3000    1.00 1971        4       4.5                   0
## 32      2800     2810    0.20 1959        4       4.0                   1
## 33      3275     2800    1.08 1956        4       3.0                   0
## 34      4500     3910    1.01 1961        5       3.0                   0
## 35      4400     3650    0.77 1966        5       5.0                   1
## 36      2375     2550    0.46 1960        4       2.5                   1
## 37      3000     3150    3.74 1959        4       3.5                   0
```

```
## 38      6000     4200    2.09 1957        4       3.0                  0
## 39      1925     1855    0.33 1972        3       3.0                  1
## 40     10500     3850    1.08 1962        3       4.0                  0
## 41      3375     2400    2.56 1955        4       2.0                  0
## 42      4750     4270    0.92 1998        4       3.5                  0
## 43      4001     3860    1.02 1987        4       4.0                  0
## 44      3600     3430    1.01 1964        4       3.0                  0
## 45      2870     3020    0.21 1955        5       3.0                  1
## 46      5900     3700    3.31 1955        4       3.0                  0
## 47      2300     2280    0.34 1972        3       2.0                  1
## 48      4194     3166    2.24 1972        2       2.5                  0
## 49      2300     2170    0.22 1950        3       2.5                  1
##    Neighborhood_CentralPV
## 1                       0
## 2                       0
## 3                       0
## 4                       0
## 5                       1
## 6                       0
## 7                       1
## 8                       0
## 9                       0
## 10                      1
## 11                      0
## 12                      0
## 13                      1
## 14                      0
## 15                      0
## 16                      0
## 17                      0
## 18                      1
## 19                      1
## 20                      1
## 21                      0
## 22                      1
## 23                      0
## 24                      1
## 25                      0
## 26                      1
## 27                      0
## 28                      1
## 29                      0
## 30                      1
## 31                      1
## 32                      0
## 33                      1
## 34                      0
## 35                      0
## 36                      0
## 37                      1
## 38                      0
## 39                      0
## 40                      1
## 41                      0
```

```
## 42                      1
## 43                      1
## 44                      0
## 45                      0
## 46                      0
## 47                      0
## 48                      1
## 49                      0
```

We see that it was loaded perfectly. There are no NaNs. A perfect dataset.

We check the summary statistics for the datasets

```
summary(RealData.df)
```

```
##    Price1000        SizeSqFt       LotAcre            Year
##  Min.   : 1925   Min.   :1660   Min.   : 0.200   Min.   :1950
##  1st Qu.: 2870   1st Qu.:2640   1st Qu.: 0.340   1st Qu.:1956
##  Median : 4025   Median :3150   Median : 1.020   Median :1959
##  Mean   : 5069   Mean   :3446   Mean   : 1.922   Mean   :1969
##  3rd Qu.: 5695   3rd Qu.:3900   3rd Qu.: 2.060   3rd Qu.:1972
##  Max.   :16988   Max.   :8333   Max.   :17.990   Max.   :2015
##     Bedrooms       Bathrooms    Neigborhood_Ladera Neighborhood_CentralPV
##  Min.   :2.000   Min.   :2.00   Min.   :0.0000     Min.   :0.0000
##  1st Qu.:3.000   1st Qu.:3.00   1st Qu.:0.0000     1st Qu.:0.0000
##  Median :4.000   Median :3.00   Median :0.0000     Median :0.0000
##  Mean   :3.918   Mean   :3.51   Mean   :0.3265     Mean   :0.3878
##  3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:1.0000     3rd Qu.:1.0000
##  Max.   :6.000   Max.   :8.00   Max.   :1.0000     Max.   :1.0000
```

Let's check for outliers. Here we check the mean and median for the price variable.

```
mean(RealData.df$Price1000)
```

```
## [1] 5069.367
```

```
median(RealData.df$Price1000)
```

```
## [1] 4025
```

Since the median is lower then the mean we see that there are going to be some outliers.

Let's check the minimum and maximum of the price variable.

```
min(RealData.df$Price1000); max(RealData.df$Price1000)
```

```
## [1] 1925
```

```
## [1] 16988
```

Now we check the lot size and see if this could be a determining factor

```
min(RealData.df$LotAcre); max(RealData.df$LotAcre)
```

```
## [1] 0.2
```

```
## [1] 17.99
```

We look at the year things were built.

```
min(RealData.df$Year); max(RealData.df$Year)
```

```
## [1] 1950
```

```
## [1] 2015
```

Let's check the number of bedrooms.

```
min(RealData.df$Bedrooms); max(RealData.df$Bedrooms)
```

```
## [1] 2
```

```
## [1] 6
```

And lastly the number of bathrooms

```
min(RealData.df$Bathrooms); max(RealData.df$Bathrooms)
```

```
## [1] 2
```

```
## [1] 8
```