ECE 625 Final Project

Gonzalo Rubio and Breno Bahia

Load data

Load data and check fields:

```
rm(list=ls()) # Clear memory

# Provided data
EdData = read.csv("./EdmontonRealEstateData.csv/EdmontonRealEstateData.csv", header=T,na.strings="?")
names(EdData)
```

```
##  [1] "taxroll_number"            "landuse_description"
##  [3] "building_name"             "market_building_class"
##  [5] "property_type"             "effective_build_year"
##  [7] "net_area"                  "basement_finished"
##  [9] "has_garage"                "has_fireplace"
## [11] "assessed_value"            "house_suit"
## [13] "house_number"              "house_suff"
## [15] "street_name"               "postal_code"
## [17] "city"                      "full_address"
## [19] "neighbourhood"             "fully_taxable"
## [21] "fully_complete"            "lot_size"
## [23] "building_count"            "build_year_mbc"
## [25] "walkout_basement"          "air_conditioning"
## [27] "valuation_group"           "display_type"
## [29] "site_coverage"             "tot_gross_area_description"
## [31] "geometry"                  "result_code"
## [33] "result_message"            "result_description"
## [35] "lon"                       "lat"
```

Data Preprocessing

Delete unwanted fields:

```
EdData = EdData[,-c(1,2,3,4,12,14,17,18,19,24,31,32,33,34)]
```

We preprocessed the data as to remove some predictors. Our of 35 initial features provided in the data, we deleted 15 of them. Not only we interpret some of these predictors as not useful (e.g., house suit and house stuff), we notice that some predictors were just combinations of others (e.g., full address combines house number, street name, postal code, and city). The deleted variable and associated reason for such action is given below

- taxroll number - used as ID only, not as data
- landuse description - could not find use yet
- building name - I wouldn't judge as important
- market building class - important but hard to handle
- house suit - empty

- house number - relative position on street, we hope to capture that in the location info
- street name - relative position on city, we hope to capture that in the location info
- postal code - relative position on city, we hope to capture that in the location info
- city - all Edmonton
- full address - relative position on province, we hope to capture that in the location info
- build year mbc - year is being used already, mbc not yet
- geometry - could identify complex geometry with corners >= 4
- result code - all same
- result message - all same
- result description - empty

Filtering

Now we preprocess the data:

```
# filter net_area < 10
EdData = EdData[EdData$net_area[] >= 10,]

# filter build count < 1
EdData = EdData[EdData$building_count[] >= 1,]

# filter assessed_value > 500
EdData = EdData[EdData$assessed_value > 500,]
```

Missing values completion

Filling missing values

```
# check missing data
apply(EdData,2,function(x) sum(is.na(x)))
```

```
##              property_type       effective_build_year
##                          0                          4
##                   net_area          basement_finished
##                          0                          0
##                 has_garage              has_fireplace
##                          0                          0
##             assessed_value               house_number
##                          0                          0
##                street_name                postal_code
##                          0                          0
##             fully_taxable            fully_complete
##                          0                          0
##                   lot_size            building_count
##                        818                          0
##           walkout_basement           air_conditioning
##                          0                          0
##            valuation_group               display_type
##                          0                          0
```

```
##              site_coverage tot_gross_area_description
##                       1253                          0
##                        lon                        lat
##                          0                          0
```

```
# assign mean year to missing years
EdData$effective_build_year[is.na(EdData$effective_build_year)] = round(mean(na.omit(EdData$effective_bu

# guesstimate missing lot size (816/818 are multi-res)
MULTIRES = na.omit(EdData[(EdData$property_type == "MULTI-RES"),])
sc = mean(MULTIRES$site_coverage)
EdData$site_coverage[is.na(EdData$lot_size)] = sc
EdData$lot_size[is.na(EdData$lot_size)] = EdData$tot_gross_area_description[is.na(EdData$lot_size)]/(sc/

# guesstimate zero lot sizes (8/12 are multi-res)
EdData$lot_size[EdData$lot_size == 0] = EdData$tot_gross_area_description[EdData$lot_size == 0]/(sc/100)

# guesstimate missing site coverage (402/435 are multi-res)
EdData$site_coverage[is.na(EdData$site_coverage)] = EdData$tot_gross_area_description[is.na(EdData$site_

# guesstimate zero site coverages (213/260 are agriculture)
AGR = EdData[(EdData$property_type == "AGRICULTURE" & EdData$site_coverage != 0),]
EdData$site_coverage[EdData$site_coverage == 0] = EdData$tot_gross_area_description[EdData$site_coverage
```

Double-check non-missing data

```
# check missing data
apply(EdData,2,function(x) sum(is.na(x)))
```

```
##              property_type       effective_build_year
##                          0                          0
##                   net_area          basement_finished
##                          0                          0
##                 has_garage              has_fireplace
##                          0                          0
##             assessed_value               house_number
##                          0                          0
##                street_name               postal_code
##                          0                          0
##              fully_taxable             fully_complete
##                          0                          0
##                   lot_size             building_count
##                          0                          0
##           walkout_basement           air_conditioning
##                          0                          0
##            valuation_group               display_type
##                          0                          0
##              site_coverage tot_gross_area_description
##                          0                          0
##                        lon                        lat
##                          0                          0
```

Double check non-zero data

```
# check zero data
colSums((EdData == 0))
```

```
##              property_type       effective_build_year
##                          0                          0
##                   net_area          basement_finished
##                          0                          0
##                 has_garage              has_fireplace
##                          0                          0
##             assessed_value               house_number
##                          0                          0
##                street_name               postal_code
##                          0                          0
##             fully_taxable             fully_complete
##                          0                          0
##                   lot_size             building_count
##                          0                          0
##           walkout_basement           air_conditioning
##                          0                          0
##            valuation_group               display_type
##                          0                          0
##              site_coverage tot_gross_area_description
##                          0                          0
##                        lon                        lat
##                          0                          0
```

summary data

```
summary(EdData)
```

```
##      property_type   effective_build_year    net_area
##  AGRICULTURE:  384   Min.   :1904         Min.   :    10.0
##  COMMERCIAL : 1120   1st Qu.:1975         1st Qu.:   108.4
##  INDUSTRIAL : 3758   Median :1986         Median :   141.0
##  MULTI-RES  : 7855   Mean   :1987         Mean   :   376.0
##  RESIDENTIAL:50516   3rd Qu.:2004         3rd Qu.:   201.8
##  URBAN      :  148   Max.   :2014         Max.   :119965.0
##
##  basement_finished has_garage  has_fireplace assessed_value
##     :    0         NO :14585   NO :25089     Min.   :     4500
##  NO :35890         Yes:49196   Yes:38692     1st Qu.:   338000
##  Yes:27891                                   Median :   422000
##                                              Mean   :   859207
##                                              3rd Qu.:   563000
##                                              Max.   :548416500
##
##   house_number                 street_name     postal_code
##  Min.   :    1   107 AVENUE NW       :  309          : 1301
##  1st Qu.: 1731   ABBOTTSFIELD ROAD NW:  290   T5T2K7 :  166
##  Median : 5867   139 AVENUE NW       :  251   T6L6J5 :  124
##  Mean   : 7134   HUNTINGTON HILL NW  :  250   T5C3C4 :  120
##  3rd Qu.:11715   SADDLEBACK ROAD NW  :  239   T6H5Y6 :  110
##  Max.   :25820   178 STREET NW       :  211   T5T1M6 :  108
##                  (Other)             :62231   (Other):61852
##  fully_taxable fully_complete    lot_size         building_count
##  NO :  532        :    2      Min.   :      0.2   Min.   : 1.000
##  Yes:63249     NO : 1043      1st Qu.:    394.0   1st Qu.: 1.000
##                Yes:62736      Median :    550.7   Median : 1.000
```

```
##                               Mean   :   2539.3   Mean   : 1.069
##                               3rd Qu.:    688.4   3rd Qu.: 1.000
##                               Max.   :1465717.1   Max.   :84.000
##
##  walkout_basement air_conditioning          valuation_group
##  NO :61194        NO :57889        RESIDENTIAL SOUTH :18051
##  Yes: 2587        Yes: 5892        RESIDENTIAL NORTH :12489
##                                    RESIDENTIAL RIVVAL:10135
##                                    RESIDENTIAL WC    : 9882
##                                    RES CONDO         : 6553
##                                    INDUSTRIAL        : 2186
##                                    (Other)           : 4485
##  display_type   site_coverage    tot_gross_area_description
##  NONRES: 5100   Min.   :  0.00   Min.   :    10.03
##  RES   :58681   1st Qu.: 19.00   1st Qu.:   145.29
##                 Median : 26.00   Median :   182.54
##                 Mean   : 26.76   Mean   :   445.40
##                 3rd Qu.: 32.00   3rd Qu.:   248.20
##                 Max.   :999.00   Max.   :201260.83
##
##       lon              lat
##  Min.   :-113.7   Min.   :53.40
##  1st Qu.:-113.6   1st Qu.:53.46
##  Median :-113.5   Median :53.50
##  Mean   :-113.5   Mean   :53.52
##  3rd Qu.:-113.4   3rd Qu.:53.58
##  Max.   :-113.3   Max.   :53.72
##
```

We set site coverage to be upper limited by 100 and put the values to percentage

```
EdData$site_coverage[EdData$site_coverage > 100] = 100
EdData$site_coverage = EdData$site_coverage/100
```

Categorical variables

Now we work on the levels categorical variables

```
# Delete non-occurring factors and change "NO" to "No"
# basement
EdData$basement_finished = factor(EdData$basement_finished)
levels(EdData$basement_finished)[1] = "No"

# we set empty fully complete to "Yes" as per the majority vote for the classes for MULTI-RES and INDUS
EdData$fully_complete[EdData$fully_complete == ""] = "Yes"
levels(EdData$fully_complete)[2] = "No"
EdData$fully_complete = factor(EdData$fully_complete)

# garage
levels(EdData$has_garage)[1] = "No"

# fireplace
levels(EdData$has_fireplace)[1] = "No"
```

```r
# wo_basement
levels(EdData$walkout_basement)[1] = "No"

# ac
levels(EdData$air_conditioning)[1] = "No"
```

Check

```r
summary(EdData)
```

```
##      property_type   effective_build_year    net_area
##  AGRICULTURE:  384   Min.   :1904         Min.   :    10.0
##  COMMERCIAL : 1120   1st Qu.:1975         1st Qu.:   108.4
##  INDUSTRIAL : 3758   Median :1986         Median :   141.0
##  MULTI-RES  : 7855   Mean   :1987         Mean   :   376.0
##  RESIDENTIAL:50516   3rd Qu.:2004         3rd Qu.:   201.8
##  URBAN      :  148   Max.   :2014         Max.   :119965.0
##
##  basement_finished has_garage  has_fireplace assessed_value
##  No :35890         No :14585   No :25089     Min.   :     4500
##  Yes:27891         Yes:49196   Yes:38692     1st Qu.:   338000
##                                              Median :   422000
##                                              Mean   :   859207
##                                              3rd Qu.:   563000
##                                              Max.   :548416500
##
##   house_number                 street_name     postal_code
##  Min.   :    1   107 AVENUE NW       :  309          : 1301
##  1st Qu.: 1731   ABBOTTSFIELD ROAD NW:  290   T5T2K7 :  166
##  Median : 5867   139 AVENUE NW       :  251   T6L6J5 :  124
##  Mean   : 7134   HUNTINGTON HILL NW  :  250   T5C3C4 :  120
##  3rd Qu.:11715   SADDLEBACK ROAD NW  :  239   T6H5Y6 :  110
##  Max.   :25820   178 STREET NW       :  211   T5T1M6 :  108
##                  (Other)             :62231   (Other):61852
##  fully_taxable fully_complete    lot_size         building_count
##  NO :  532     No : 1043      Min.   :      0.2   Min.   : 1.000
##  Yes:63249     Yes:62738      1st Qu.:    394.0   1st Qu.: 1.000
##                               Median :    550.7   Median : 1.000
##                               Mean   :   2539.3   Mean   : 1.069
##                               3rd Qu.:    688.4   3rd Qu.: 1.000
##                               Max.   :1465717.1   Max.   :84.000
##
##  walkout_basement air_conditioning           valuation_group
##  No :61194        No :57889        RESIDENTIAL SOUTH :18051
##  Yes: 2587        Yes: 5892        RESIDENTIAL NORTH :12489
##                                    RESIDENTIAL RIVVAL:10135
##                                    RESIDENTIAL WC    : 9882
##                                    RES CONDO         : 6553
##                                    INDUSTRIAL        : 2186
##                                    (Other)           : 4485
##  display_type   site_coverage       tot_gross_area_description
##  NONRES: 5100   Min.   :0.0000005   Min.   :    10.03
##  RES   :58681   1st Qu.:0.1900000   1st Qu.:   145.29
##                 Median :0.2600000   Median :   182.54
```

```
##                 Mean   :0.2643572   Mean   :   445.40
##                 3rd Qu.:0.3200000   3rd Qu.:   248.20
##                 Max.   :1.0000000   Max.   :201260.83
##
##       lon              lat
##  Min.   :-113.7   Min.   :53.40
##  1st Qu.:-113.6   1st Qu.:53.46
##  Median :-113.5   Median :53.50
##  Mean   :-113.5   Mean   :53.52
##  3rd Qu.:-113.4   3rd Qu.:53.58
##  Max.   :-113.3   Max.   :53.72
##
```

Numerical variables

Get bulding age (we overwrite year with 2015 - year)

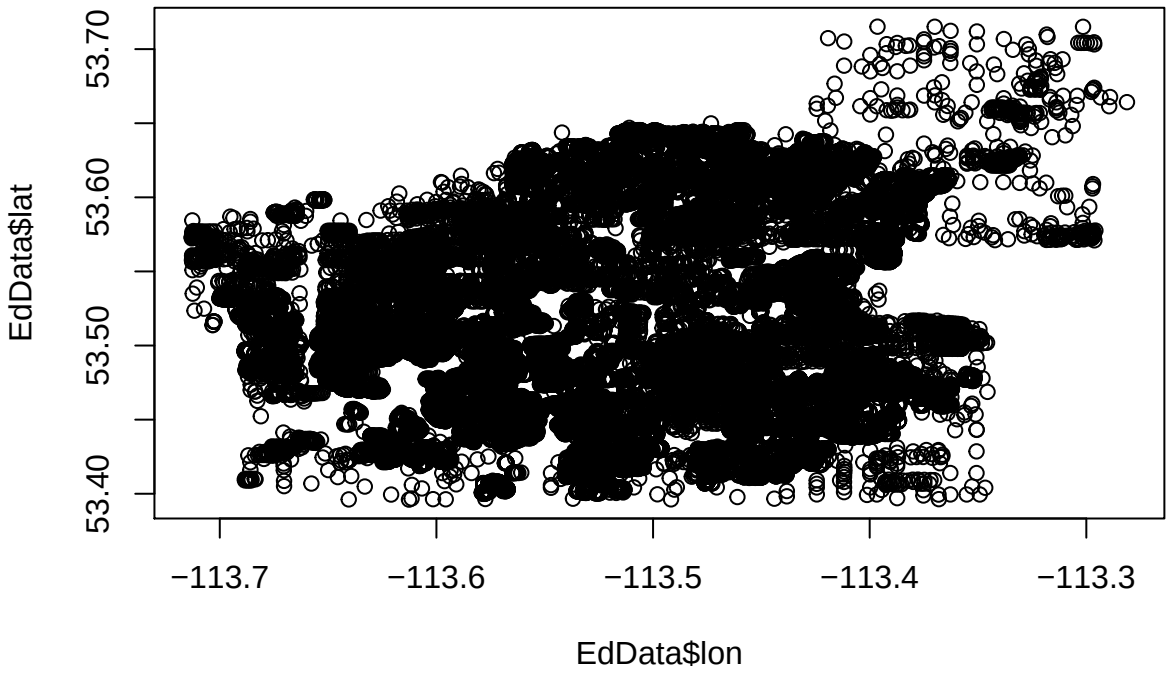
```
EdData$effective_build_year = 2015 - EdData$effective_build_year
```

Divide area into North, South, East, West based on the mean latitude and longitude

```
# mid points
midLat = mean(EdData$lat)
midLon = mean(EdData$lon)
# categorical variable as factors
EdData$quadrant = as.factor(ifelse(EdData$lon > midLon&EdData$lat > midLat, "NE",
                      ifelse(EdData$lon > midLon&EdData$lat < midLat, "NW",
                      ifelse(EdData$lon < midLon&EdData$lat > midLat, "SE", "SW"))))
# data frame
EdData = data.frame(EdData)
```

Plots

```
plot(EdData$lon,EdData$lat)
```


EdData$lat
53.40
53.50
53.60
53.70
−113.7
−113.6
−113.5
−113.4
−113.3
EdData$lon