Assignment 2

```r
dat <- fread("./data/properties_2016.csv")
```

```
##
Read 0.0% of 2985217 rows
```

```
## Warning in fread("./data/properties_2016.csv"): Bumped column 50 to type
## character on data row 10354, field contains 'true'. Coercing previously
## read values in this column from logical, integer or numeric back to
## character which may not be lossless; e.g., if '00' and '000' occurred
## before they will now be just '0', and there may be inconsistencies with
## treatment of ',,' and ',NA,' too (if they occurred in this column before
## the bump). If this matters please rerun and set 'colClasses' to 'character'
## for this column. Please note that column type detection uses a sample of
## 1,000 rows (100 rows at 10 points) so hopefully this message should be very
## rare. If reporting to datatable-help, please rerun and include the output
## from verbose=TRUE.
```

```
##
Read 16.7% of 2985217 rows
Read 33.8% of 2985217 rows
Read 50.6% of 2985217 rows
Read 67.3% of 2985217 rows
Read 84.4% of 2985217 rows
Read 2985217 rows and 58 (of 58) columns from 0.604 GB file in 00:00:09
```

Explain why it is or why it is not a big data problem.

Why is it an analytics problem?

How many data attributes are there?

As can be seen from the result of `colnames(dat)`, we know that the data set contains 58 attributes.

```r
colnames(dat)
```

```
##  [1] "parcelid"                     "airconditioningtypeid"
##  [3] "architecturalstyletypeid"     "basementsqft"
##  [5] "bathroomcnt"                  "bedroomcnt"
##  [7] "buildingclasstypeid"          "buildingqualitytypeid"
##  [9] "calculatedbathnbr"            "decktypeid"
## [11] "finishedfloor1squarefeet"     "calculatedfinishedsquarefeet"
## [13] "finishedsquarefeet12"         "finishedsquarefeet13"
## [15] "finishedsquarefeet15"         "finishedsquarefeet50"
## [17] "finishedsquarefeet6"          "fips"
## [19] "fireplacecnt"                 "fullbathcnt"
## [21] "garagecarcnt"                 "garagetotalsqft"
## [23] "hashottuborspa"               "heatingorsystemtypeid"
## [25] "latitude"                     "longitude"
## [27] "lotsizesquarefeet"            "poolcnt"
## [29] "poolsizesum"                  "pooltypeid10"
## [31] "pooltypeid2"                  "pooltypeid7"
## [33] "propertycountylandusecode"    "propertylandusetypeid"
## [35] "propertyzoningdesc"           "rawcensustractandblock"
## [37] "regionidcity"                 "regionidcounty"
## [39] "regionidneighborhood"         "regionidzip"
## [41] "roomcnt"                      "storytypeid"
## [43] "threequarterbathnbr"          "typeconstructiontypeid"
## [45] "unitcnt"                      "yardbuildingsqft17"
## [47] "yardbuildingsqft26"           "yearbuilt"
## [49] "numberofstories"              "fireplaceflag"
## [51] "structuretaxvaluedollarcnt"   "taxvaluedollarcnt"
## [53] "assessmentyear"               "landtaxvaluedollarcnt"
## [55] "taxamount"                    "taxdelinquencyflag"
## [57] "taxdelinquencyyear"           "censustractandblock"
```

Identify the type of the 15 attributes you find most relevant in this context.

Determine whether the task refers to a supervised or unsupervised learning problem

Find out what the standard analysis algorithms are for this analytics problem.

Provide descriptive summaries of the sample data

```r
summary(dat)
```

```
##     parcelid         airconditioningtypeid architecturalstyletypeid
##  Min.   : 10711725   Min.   : 1.0          Min.   : 2.0
##  1st Qu.: 11643707   1st Qu.: 1.0          1st Qu.: 7.0
##  Median : 12545094   Median : 1.0          Median : 7.0
##  Mean   : 13325858   Mean   : 1.9          Mean   : 7.2
##  3rd Qu.: 14097122   3rd Qu.: 1.0          3rd Qu.: 7.0
##  Max.   :169601949   Max.   :13.0          Max.   :27.0
##                      NA's   :2173698       NA's   :2979156
##   basementsqft      bathroomcnt       bedroomcnt     buildingclasstypeid
##  Min.   :  20.0    Min.   : 0.000   Min.   : 0.000   Min.   :1.0
##  1st Qu.: 272.0    1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:3.0
##  Median : 534.0    Median : 2.000   Median : 3.000   Median :4.0
##  Mean   : 646.9    Mean   : 2.209   Mean   : 3.089   Mean   :3.7
##  3rd Qu.: 847.2    3rd Qu.: 3.000   3rd Qu.: 4.000   3rd Qu.:4.0
##  Max.   :8516.0    Max.   :20.000   Max.   :20.000   Max.   :5.0
##  NA's   :2983589   NA's   :11462    NA's   :11450    NA's   :2972588
##  buildingqualitytypeid calculatedbathnbr   decktypeid
##  Min.   : 1.0          Min.   : 1.0      Min.   :66
##  1st Qu.: 4.0          1st Qu.: 2.0      1st Qu.:66
##  Median : 7.0          Median : 2.0      Median :66
##  Mean   : 5.8          Mean   : 2.3      Mean   :66
##  3rd Qu.: 7.0          3rd Qu.: 3.0      3rd Qu.:66
##  Max.   :12.0          Max.   :20.0      Max.   :66
##  NA's   :1046729       NA's   :128912    NA's   :2968121
##  finishedfloor1squarefeet calculatedfinishedsquarefeet
##  Min.   :    3            Min.   :     1
##  1st Qu.: 1012            1st Qu.:  1213
##  Median : 1283            Median :  1572
##  Mean   : 1381            Mean   :  1827
##  3rd Qu.: 1615            3rd Qu.:  2136
##  Max.   :31303            Max.   :952576
##  NA's   :2782500          NA's   :55565
##  finishedsquarefeet12 finishedsquarefeet13 finishedsquarefeet15
##  Min.   :     1       Min.   : 120         Min.   :   112
##  1st Qu.:  1196       1st Qu.: 960         1st Qu.:  1694
##  Median :  1539       Median :1296         Median :  2172
##  Mean   :  1760       Mean   :1179         Mean   :  2739
##  3rd Qu.:  2070       3rd Qu.:1440         3rd Qu.:  2976
##  Max.   :290345       Max.   :2688         Max.   :820242
##  NA's   :276033       NA's   :2977545      NA's   :2794419
##  finishedsquarefeet50 finishedsquarefeet6      fips
##  Min.   :    3        Min.   :   117      Min.   :6037
##  1st Qu.: 1013        1st Qu.:  1079      1st Qu.:6037
##  Median : 1284        Median :  1992      Median :6037
##  Mean   : 1389        Mean   :  2414      Mean   :6048
##  3rd Qu.: 1618        3rd Qu.:  3366      3rd Qu.:6059
##  Max.   :31303        Max.   :952576      Max.   :6111
##  NA's   :2782500      NA's   :2963216     NA's   :11437
##   fireplacecnt      fullbathcnt      garagecarcnt     garagetotalsqft
##  Min.   :1.0       Min.   : 1.00    Min.   : 0.0      Min.   :   0.0
##  1st Qu.:1.0       1st Qu.: 2.00    1st Qu.: 2.0      1st Qu.: 324.0
##  Median :1.0       Median : 2.00    Median : 2.0      Median : 441.0
##  Mean   :1.2       Mean   : 2.24    Mean   : 1.8      Mean   : 383.8
##  3rd Qu.:1.0       3rd Qu.: 3.00    3rd Qu.: 2.0      3rd Qu.: 494.0
##  Max.   :9.0       Max.   :20.00    Max.   :25.0      Max.   :7749.0
##  NA's   :2672580   NA's   :128912   NA's   :2101950   NA's   :2101950
##  hashottuborspa     heatingorsystemtypeid    latitude
##  Length:2985217     Min.   : 1           Min.   :33324388
##  Class :character   1st Qu.: 2           1st Qu.:33827685
##  Mode  :character   Median : 2           Median :34008249
##                     Mean   : 4           Mean   :34001469
##                     3rd Qu.: 7           3rd Qu.:34161860
##                     Max.   :24           Max.   :34819650
##                     NA's   :1178816      NA's   :11437
##    longitude          lotsizesquarefeet      poolcnt
##  Min.   :-119475780   Min.   :      100   Min.   :1
##  1st Qu.:-118392983   1st Qu.:     5688   1st Qu.:1
##  Median :-118172540   Median :     7000   Median :1
##  Mean   :-118201934   Mean   :    22823   Mean   :1
##  3rd Qu.:-117949468   3rd Qu.:     9898   3rd Qu.:1
##  Max.   :-117554316   Max.   :328263808   Max.   :1
##  NA's   :11437        NA's   :276099      NA's   :2467683
##   poolsizesum       pooltypeid10      pooltypeid2       pooltypeid7
##  Min.   :   19.0   Min.   :1         Min.   :1         Min.   :1
##  1st Qu.:  430.0   1st Qu.:1         1st Qu.:1         1st Qu.:1
##  Median :  495.0   Median :1         Median :1         Median :1
##  Mean   :  519.7   Mean   :1         Mean   :1         Mean   :1
##  3rd Qu.:  594.0   3rd Qu.:1         3rd Qu.:1         3rd Qu.:1
##  Max.   :17410.0   Max.   :1         Max.   :1         Max.   :1
##  NA's   :2957257   NA's   :2948278   NA's   :2953142   NA's   :2499758
##  propertycountylandusecode propertylandusetypeid propertyzoningdesc
##  Length:2985217            Min.   : 31           Length:2985217
##  Class :character          1st Qu.:261           Class :character
##  Mode  :character          Median :261           Mode  :character
##                            Mean   :260
##                            3rd Qu.:261
##                            Max.   :275
##                            NA's   :11437
##  rawcensustractandblock  regionidcity    regionidcounty
##  Min.   :60371011       Min.   :  3491   Min.   :1286
##  1st Qu.:60373203       1st Qu.: 12447   1st Qu.:2061
##  Median :60375712       Median : 25218   Median :3101
##  Mean   :60483450       Mean   : 34993   Mean   :2570
##  3rd Qu.:60590423       3rd Qu.: 45457   3rd Qu.:3101
##  Max.   :61110091       Max.   :396556   Max.   :3101
##  NA's   :11437          NA's   :62845    NA's   :11437
##  regionidneighborhood  regionidzip        roomcnt        storytypeid
##  Min.   :  6952       Min.   : 95982   Min.   : 0.000   Min.   :7
##  1st Qu.: 46736       1st Qu.: 96180   1st Qu.: 0.000   1st Qu.:7
##  Median :118920       Median : 96377   Median : 0.000   Median :7
##  Mean   :193476       Mean   : 96553   Mean   : 1.475   Mean   :7
##  3rd Qu.:274800       3rd Qu.: 96974   3rd Qu.: 0.000   3rd Qu.:7
##  Max.   :764167       Max.   :399675   Max.   :96.000   Max.   :7
##  NA's   :1828815      NA's   :13980    NA's   :11475    NA's   :2983593
##  threequarterbathnbr typeconstructiontypeid    unitcnt
##  Min.   :1           Min.   : 4             Min.   :  1.0
##  1st Qu.:1           1st Qu.: 6             1st Qu.:  1.0
##  Median :1           Median : 6             Median :  1.0
##  Mean   :1           Mean   : 6             Mean   :  1.2
##  3rd Qu.:1           3rd Qu.: 6             3rd Qu.:  1.0
##  Max.   :7           Max.   :13             Max.   :997.0
##  NA's   :2673586     NA's   :2978470        NA's   :1007727
##  yardbuildingsqft17 yardbuildingsqft26   yearbuilt     numberofstories
##  Min.   :  10.0     Min.   :  10.0     Min.   :1801    Min.   : 1.0
##  1st Qu.: 190.0     1st Qu.:  96.0     1st Qu.:1950    1st Qu.: 1.0
##  Median : 270.0     Median : 168.0     Median :1963    Median : 1.0
##  Mean   : 319.8     Mean   : 278.3     Mean   :1964    Mean   : 1.4
##  3rd Qu.: 390.0     3rd Qu.: 320.0     3rd Qu.:1981    3rd Qu.: 2.0
##  Max.   :7983.0     Max.   :6141.0     Max.   :2015    Max.   :41.0
##  NA's   :2904862    NA's   :2982570    NA's   :59928   NA's   :2303148
##  fireplaceflag      structuretaxvaluedollarcnt taxvaluedollarcnt
##  Length:2985217     Min.   :        1          Min.   :        1
##  Class :character   1st Qu.:    74800          1st Qu.:   179675
##  Mode  :character   Median :   122590          Median :   306086
##                     Mean   :   170884          Mean   :   420479
##                     3rd Qu.:   196889          3rd Qu.:   488000
##                     Max.   :251486000          Max.   :282786000
##                      NA's   :54982              NA's   :42550
##  assessmentyear  landtaxvaluedollarcnt   taxamount
##  Min.   :2000    Min.   :       1      Min.   :      1
##  1st Qu.:2015    1st Qu.:   74836      1st Qu.:   2461
##  Median :2015    Median :  167042      Median :   3992
##  Mean   :2015    Mean   :  252478      Mean   :   5378
##  3rd Qu.:2015    3rd Qu.:  306918      3rd Qu.:   6201
##  Max.   :2016    Max.   :90246219      Max.   :3458861
##  NA's   :11439   NA's   :67733         NA's   :31250
##  taxdelinquencyflag taxdelinquencyyear censustractandblock
##  Length:2985217     Min.   : 0.0       Min.   :             -1
##  Class :character   1st Qu.:14.0       1st Qu.: 60374002041015
##  Mode  :character   Median :14.0       Median : 60375715022011
##                     Mean   :13.9       Mean   : 60484312212563
##                     3rd Qu.:15.0       3rd Qu.: 60590423191014
##                     Max.   :99.0       Max.   :483030105084015
##                     NA's   :2928753    NA's   :          75126
```

How are the missings distributed?