Lab 04 - Describing Data

Francisco Santamarina

September 27, 2016

Load the necessary packages and dataset.

```
##load data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pander)
dat <- read.csv( "https://raw.githubusercontent.com/lecy/maps-in-R/master/Data/syr_parcels.csv" )
```

Question 1

What class is the dataset? "data.frame"

```
class(dat)
```

```
## [1] "data.frame"
```

How many rows of data are there? 41502 rows of unique data

```
nrow(dat)
```

```
## [1] 41502
```

How many variables? 64 variables

```
length(dat)
```

```
## [1] 64
```

How would you check the class of a variables in the dataset?

```
names(dat)
```

```
##  [1] "TAX_ID"     "PRINTKEY"   "FRONTFEET"  "DEPTH"      "SqFt"
##  [6] "Acres"      "Sec_Block"  "TAX_ID_1"   "SURA"       "Quad"
## [11] "Nhood"      "TNT_NAME"   "Special_Nh" "Assessment" "CensusTrac"
## [16] "CC_Dist"    "COUNTY_LEG" "SEIZB"      "Owner"      "LUCODE"
## [21] "LandUse"    "Units"      "AmtDelinqu" "Totint"     "TaxYrsDeli"
## [26] "StNum"      "StName"     "AS400_OCV"  "IPS_OCV"    "Condition"
## [31] "AssessedLa" "AssessedVa" "VacantBuil" "DVDATE"     "CityTaxabl"
## [36] "STARS"      "STARC"      "STAR"       "Owner2"     "Add1"
## [41] "Add2"       "Add3"       "Add4"       "ZIP"        "ZIP2"
## [46] "WaterServi" "YearBuilt"  "SALES"      "PNUMBR"     "OverdueWat"
## [51] "WARD"       "SBL"        "CountyTXBL" "SchoolTXBL" "Bankruptcy"
## [56] "TOTSYR"     "TOTONO"     "INTSYR"     "INTONO"     "TaxTrust"
## [61] "SENIOR_EXE" "VET_EXEMPT" "Redemption" "Round"
```

```
class(dat$TAX_ID)
```

```
## [1] "numeric"
```

Question 2

2. Convert your dataset into a tibble using the tbl_df() function. What is the class of the dataset now? How can you check rows, columns, and class of variables for the tibble?

```
tbl.dat <- tbl_df(dat)
```

What is the class of the dataset now? It is a tbl and inherits the class data.frame

```
class(tbl.dat)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
typeof(tbl.dat)
```

```
## [1] "list"
```

How can you check rows, columns, and class of variables for the tibble?

```
#Rows/Observations, Columns/Variables, and class:
glimpse(tbl.dat)
```

```
## Observations: 41,502
## Variables: 64
## $ TAX_ID     <dbl> 3.115001e+25, 3.115001e+25, 3.115001e+25, 3.115001e...
## $ PRINTKEY   <fctr> 065.1-03-01.0, 065.1-03-02.0, 064.-13-15.0, 064.-1...
## $ FRONTFEET  <dbl> 67.20, 104.80, 82.87, 65.00, 65.00, 65.00, 65.00, 6...
## $ DEPTH      <dbl> 50.00, 46.50, 168.28, 160.89, 160.89, 160.89, 160.8...
```

```
## $ SqFt       <dbl> 2149.182, 6370.403, 12910.006, 10322.661, 10457.843...
## $ Acres      <dbl> 0.04933843, 0.14624434, 0.29637297, 0.23697569, 0.2...
## $ Sec_Block  <dbl> 3.115001e+15, 3.115001e+15, 3.115001e+15, 3.115001e...
## $ TAX_ID_1   <dbl> 3.115001e+25, 3.115001e+25, 3.115001e+25, 3.115001e...
## $ SURA       <fctr> N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N,...
## $ Quad       <fctr> SW, SW, SE, SE, SE, SE, SE, SE, SW, SE, SE, SW, SE...
## $ Nhood      <fctr> South Valley, South Valley, South Valley, South Va...
## $ TNT_NAME   <fctr> Valley, Valley, Valley, Valley, Valley, Valley, Va...
## $ Special_Nh <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ Assessment <int> 39, 39, 37, 37, 37, 37, 37, 37, 39, 37, 37, 37, 37,...
## $ CensusTrac <fctr> 60, 60, 61.03, 61.03, 61.03, 61.03, 61.03, 61.03, ...
## $ CC_Dist    <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ COUNTY_LEG <int> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15,...
## $ SEIZB      <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, Y, NA,...
## $ Owner      <fctr> CLARMIN BUILDERS ONON COR, JOHNSTON LEE R, CHRISTO...
## $ LUCODE     <int> 312, 210, 210, 210, 210, 210, 210, 210, 311, 210, 2...
## $ LandUse    <fctr> Vacant Land, Single Family, Single Family, Single ...
## $ Units      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ AmtDelinqu <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ Totint     <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ TaxYrsDeli <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 17, 0, 0, 0, 0, 0,...
## $ StNum      <fctr> 2655, 2635, 203, 100, 104, 108, 112, 116, 301, 120...
## $ StName     <fctr> VALLEY DR, VALLEY DR, HAYES TERR, EDNA RD & CITY L...
## $ AS400_OCV  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
## $ IPS_OCV    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ Condition  <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ AssessedLa <dbl> 475, 10800, 20200, 18000, 18000, 18000, 18000, 1800...
## $ AssessedVa <dbl> 500, 69300, 88300, 70500, 74000, 95000, 72000, 6850...
## $ VacantBuil <fctr> N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N,...
## $ DVDATE     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 990000, 0, 0, 9900...
## $ CityTaxabl <dbl> 500, 69300, 88300, 70500, 74000, 95000, 72000, 6700...
## $ STARS      <dbl> 500, 52320, 71320, 70500, 57020, 78020, 55020, 3116...
## $ STARC      <dbl> 500, 52320, 71320, 70500, 57020, 78020, 55020, 3116...
## $ STAR       <fctr> NA, Y, Y, NA, Y, Y, Y, Y, NA, Y, Y, NA, Y, Y, NA, ...
## $ Owner2     <fctr> NA, NA, CHRISTO TERRI A, NA, NA, NA, WHALEN WILLIA...
## $ Add1       <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ Add2       <fctr> NA, NA, NA, PO BOX B, NA, NA, NA, NA, NA, NA, NA, ...
## $ Add3       <fctr> 4604 BEEF ST, 2635 VALLEY DR, 112 RAMSEY AVE, NA, ...
## $ Add4       <fctr> SYRACUSE  NY, SYRACUSE  NY, SYRACUSE  NY, NEDROW  ...
## $ ZIP        <fctr> 13215, 13120, 13224, 13120, 13205, 13205, 13205, 1...
## $ ZIP2       <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ WaterServi <fctr> NA, A, A, A, A, A, A, A, NA, A, A, NA, A, A, NA, A...
## $ YearBuilt  <int> NA, 1925, 1957, 1958, 1965, 1954, 1953, 1955, NA, 1...
## $ SALES      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ PNUMBR     <int> 1393130501, 1393130500, 1437100600, 1425100900, 142...
## $ OverdueWat <dbl> 0.00, 177.79, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0...
## $ WARD       <int> 13, 13, 14, 14, 14, 14, 14, 14, 13, 14, 14, 14, 14,...
## $ SBL        <fctr> 065.1-03-01.0, 065.1-03-02.0, 064.-13-15.0, 064.-1...
## $ CountyTXBL <dbl> 500, 69300, 88300, 70500, 74000, 95000, 72000, 6700...
## $ SchoolTXBL <dbl> 500, 69300, 88300, 70500, 74000, 95000, 72000, 6700...
## $ Bankruptcy <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ TOTSYR     <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ TOTONO     <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ INTSYR     <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
```

```
## $ INTONO     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 816, 0, 0, 0, 0, 0...
## $ TaxTrust   <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ SENIOR_EXE <fctr> NA, NA, NA, NA, NA, NA, NA, Y, NA, NA, NA, NA, NA,...
## $ VET_EXEMPT <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ Redemption <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0...
## $ Round      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

Question 3

3. Create a subset of the data by retaining the following set of variables:

- Acres, LandUse, AmtDelinqu, AssessedLa, VacantBuil, YearBuilt, Nhood

```
littletbl <- select(.data=tbl.dat, Acres, LandUse, AmtDelinqu, AssessedLa, VacantBuil, YearBuilt, Nhood
```

Question 4

4. Drop the cases there the year of construction (YearBuilt) is reported as zero using the filter() function.

```
no.0.littletbl <- filter(.data = littletbl, YearBuilt != 0)
```

Question 5

5. Create a new variable that measures the assessed land value (AssessedLa) per acre.

```
alv.acre <- mutate(.data = no.0.littletbl, AssessedLa.acre = AssessedLa / Acres)
```

Question 6

6. Which neighborhood has the highest average land value per acre? Use the group_by() and summarise() functions to answer this question. Downtown, at an average of $895,080.30 per acre

```
#g.alv.acre <- group_by(alv.acre, Nhood)
#summ.g <- summarise( g.alv.acre, AvgALV.acre = mean( AssessedLa.acre))
#arrange(summ.g, desc( AvgALV.acre) )

arrange(summarise( group_by(alv.acre, Nhood), AvgALV.acre = mean( AssessedLa.acre)), desc( AvgALV.acre)
```

```
## # A tibble: 33 × 2
##              Nhood AvgALV.acre
##             <fctr>       <dbl>
## 1         Downtown    895080.3
## 2  University Hill    361726.9
## 3  Franklin Square    212634.0
## 4    Prospect Hill    146814.8
```

```
## 5        Lakefront    126197.9
## 6         Sedgwick    118333.2
## 7     Hawley-Green    117458.2
## 8     Far Westside    113103.9
## 9        Tipp Hill    106725.7
## 10        Eastwood    106283.8
## # ... with 23 more rows
```

Question 7

7. OPTIONAL: Create a new variable that splits the year of construction up by decade using the cut() function. Which decade produced single family homes with the highest assessed value (AssessedVa)?