NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

R Lesson 1 - Solutions
MSPA 401 – Introduction to Statistical Analysis

Exercises:

a) What are the measurement levels of each of the eight variables?

Ratio: PRICE, SQFT, TAX, FEATS
Interval: YEAR
Ordinal: BATHS
Nominal: NBR, CORNER
Nominal or ordinal measurements cannot be measured numerically and are referred to as categorical variables. Ratio and interval measurements are referred to as quantitative variables and can be either discrete or continuous depending on their nature.
FEATS is a count. Since there can be a zero count for FEATS, and the arithmetic difference between two counts is meaningful, FEATS is a discrete quantitative variable at the ratio level.
YEAR is an interval variable since there is no firm zero starting time. YEAR equal to zero is by definition or convention.
BATHS is not really a count. This is an instance where numbers are assigned to ordered categories. Although there can be zero baths theoretically, the ratios and differences that may be formed are not meaningful. For example, what is meaningful about the ratio between 1.5 baths and 2.5 baths? From another point of view, is the difference between 2 and 3 baths twice that of between 1.0 and 1.5 baths? Since the ratios and differences are not meaningful, BATHS is categorical at the ordinal level measurement.

b) Should any variable have its values changed to better reflect its true nature?

YEAR could be expressed in terms of house age. This would not change its nature since the age would keep changing depending on what was taken to be the present date.
BATHS could be expressed in terms of an ordered scale as long as each category in the scale had a definition.

c) For the variable PRICE, select a simple random sample of size 12 from the file. Save this sample in a vector named SRS. Print the values in SRS and compute the mean value.

```
> print(SRS)
[1] 4500.0 5250.0 2575.0 2550.0 2625.0 1875.0 3882.5 2612.5 4360.0
1915.0 2187.5 1550.0
> mean(SRS)
[1] 2990.208
```

d) For the variable PRICE, select a systematic sample of twelve observations. Start with the seventh observation and pick every 10^{th} observation thereafter (i.e. 7, 17, 27,..). You should end with the 117^{th} observation. Save the sample in a vector named SS. Print the values SRS and compute the mean value. (For picking a systematic sample refer to page 283 of Lander for an example. Use seq(from,to,by=).)

```
> print(SS)
[1] 1750.0 2347.5 3250.0 3997.5 3125.0 2950.0 4062.5 5250.0 2822.5
3325.0 5375.0 3900.0
> mean(SS)
[1] 3512.91
```

e) Examine the printed values and mean values obtained from the two sampling procedures. Do you see a difference? (Try the commands summary(SRS) and summary(SS).)

```
> summary(SRS)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1550    2119    2594    2990    4002    5250
> summary(SS)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1750    2918    3288    3513    4014    5375
```

The main difference is that the values for SS appear to be larger than for SRS.

f) Create boxplots for SRS and SS using boxplot(). How do the two samples compare?

Ranges of the data are similar, however there is a shift in the mean values.