# Final Project - Group C

*Lionel Johnnes*

*March 12, 2016*

ASSIGNMENT GUIDELINES: From a dataset of your choice, build a regression, classification or clustering method that solves a challenge you must also define. The work must include a brief stat description of the dataset aligned with the goal of the analysis, and the valuation of the different alternatives and choices made during the process. 10 pages max, no code included.

Datasets included in prior public work are not allowed. The work will be assessed considering: - how relevant and thorough is the stat description, - information that will help the reproducibility (markdown code, plots, links to datasets, etc.) - the choice of the model, the goal of the analysis, the achievements and the exposition of results. The maximum number of pages that the PDF document must have is 10. Your abilities to synthesize relevant information and communicate important results will be also considered.

DATA SET: Austin Craft Realty Condo sales in Downtown Austin, Texas for the past year (03/18/2015 - 03/11/2015)

```
##
## 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
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
## Warning: replacing previous import by 'grid::arrow' when loading 'GGally'
```

```
## Warning: replacing previous import by 'grid::unit' when loading 'GGally'
```

```
##
## Attaching package: 'GGally'
##
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
## Warning: package 'arules' was built under R version 3.2.3
```

```
## Loading required package: Matrix
##
## Attaching package: 'arules'
##
## The following objects are masked from 'package:base':
##
```

```
##     %in%, abbreviate, write
##
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
##
##     between, last
```

```
## [1] 230  24
```

```
##   MLS.Number Status Area Street.Number Street.Dir.Prefix Street.Name
## 1    5591862   Sold   DT          1800                       Lavaca
## 2    8565582   Sold   DT          1800                       Lavaca
## 3    4025784   Sold   DT          1800                       Lavaca
## 4    6765009   Sold   DT          1800                       Lavaca
## 5    6018863   Sold   DT          1212                    Guadalupe
## 6    5563577   Sold   DT          1700                       Nueces
##   Street.Type Street.Dir.Suffix Unit.Number X..Beds.Total X..Baths.Full
## 1          ST                         A-306             1             1
## 2          ST                         A-412             1             1
## 3                                       212             1             1
## 4          ST                         A-614             1             1
## 5          ST                           203             1             1
## 6          ST                           106             1             1
##   X..Baths.Half X..Living X..Stories X..Garage.Spaces Year.Built
## 1             0         1          1                1       1966
## 2             0         1          1                1       1966
## 3             0         1          1                1       1966
## 4             0         1          1                1       1966
## 5             0         1          1                1       1973
## 6             0         1          3                1       1974
##   Sqft.Total LP.SqFt List.Price  S...SF Sold.Lease.Price Sold.Lease.Date
## 1        608  245.89   $149,500 $234.38         $142,500      12-14-2015
## 2        594  222.22   $132,000 $240.74         $143,000      05-15-2015
## 3        594  244.11   $144,999 $244.11         $145,000      09-02-2015
## 4        632  237.33   $149,995 $237.33         $149,995      07-10-2015
## 5        422     391   $165,000 $386.26         $163,000      01-28-2016
## 6        488  348.16   $169,900 $336.07         $164,000      06-05-2015
##   ADOM CDOM
## 1    6    6
## 2   12   12
## 3   14   14
## 4    1    1
## 5    7    7
## 6   10   10
```

The first variable, MLS.Number, is a unique identifier and can be removed as it does not aid in predictions. In addition, every observation shares one value for the following variables: Status (Sold), Area (DT) and Street.Dir.Suffix (except for 1). Again, we remove these variables from the data set.

We then create several new variables that could be useful in predictions or interesting to explore: 1. age (Numeric): How old is the home? 2. relisted (Boolean): Was the home taken off market at some point? 3. soldMonth (Categorical): In what month was it sold? 4. listMonth (Categorical): In what month was it listed?

5. soldDay (Categorical): On what day of the week was it sold? 6. Floor (Numeric): On what floor of the building is it? # we need to be careful with floor, since we may be saying units on the 12th floor are actually on the 1st. 7. Price.Diff (Numeric): How did sale price differ from list price? 8. condoSize (Categorical): What type (size: S,M,L) condo is it? 9. ccilisted (Numeric): What was the consumer confidence index the month the condo listing was posted? 10. ccisold (Numeric): What was the consumer confidence index the month the condo was sold?

```
data <- data[,-c(1,2,3,8)]

# Create new variables
data$age <- 2016 - data$Year.Built

data$relisted <- data$CDOM - data$ADOM
data$relisted[data$relisted>0] <- 1

#It was all because of using "/" instead of "-"
soldDates <- as.Date(data$Sold.Lease.Date, format = "%m-%d-%Y")
data$soldMonth <- format(as.Date(soldDates), "%m")

listDate <- soldDates-data$ADOM
data$listMonth <- format(as.Date(listDate), "%m")

data$soldDay <- weekdays(soldDates)

data$Unit.Number <- gsub('[^a-zA-Z0-9.]', '', data$Unit.Number)
data$floor <- stri_extract_first(data$Unit.Number, regex = "\\d")
data$floor <- as.numeric(data$floor)

#removing all sorts of characters (dollar signs in this case) and then changing as numeric
data$Sold.Lease.Price <- gsub('[^a-zA-Z0-9.]', '', data$Sold.Lease.Price)
data$List.Price <- gsub('[^a-zA-Z0-9.]', '', data$List.Price)
data$Price.Diff <- as.numeric(data$Sold.Lease.Price) - as.numeric(data$List.Price)

#Condo size as discrete
data$condoSize <- discretize(data$Sqft.Total, "cluster", categories = 3, labels=c("small", "medium", "la

soldDates <- as.Date(data$Sold.Lease.Date, format = "%m-%d-%Y")
data$soldMonth <- format(as.Date(soldDates), "%m")


#ccidate <- as.Date(cci$TIME, format = "%Y-%m")

#data$ccilisted

#data$ccisold
```