An Analysis on Philadelphia Real-estate

Introduction to Data Mining (DSS 660) Final Project

**Avinash Reddy Periyavaram**

Start Date - 9/30/2018

Knit Date - 10/20/2018

# Abstract

In this data analysis project, I have chosen a data set of 391 houses in Philadelphia (randomly selected from adds posted between August 2016 to March 2017), which are scraped from website zillow.com.

The project is performed using “R 3.5.1” statistical programming and “R studio” development environment.

The analysis has been divided into five parts for better understanding of the flow.

Part 1 has the information about loading the required packages

Part 2 has the information about loading the data to the R environment,

Part 3 has the process of cleaning the data

Part 4 is an overview of the data with basic descriptive statistics

Part 5 has all the data analysis.

## Conditions for Random Sampling

1. **Independence**: This data set is scraped from Zillow website. We are fine in assuming that the records are independent.
2. **Sample Size**: As the samples are obtained without replacement, we must ensure they represent less than 10% of the population. The 391 records we use for this analysis is indeed less than 10% of the total houses in Philadelphia.

# Objectives -

This project has four differnt objectives to be addressed

Predicting a house price is the first and important objective of the analysis, apart from building a predective model for price estimation it is important to find which features in the data are affecting the price of a house in Philly.  
   
 Second objective is to test wheather the mean price of differnt house types are similar or atleast one of the mean price is different. This analysis is helpful to find which types of houses have a similar a similar price range, he house types with differnt mean prices and their differences between their price ranges  
   
 In reality when I visit center city, I see mostly condos and townhouse, but in the place I live (about 10 Miles from center city) I see many single family big houses. I am interested to test weather the location matters which type of house is constructed.  
   
 Final objective is to perform an exproitory analysis in finding which houses can be grouped into a category (Cluster Analysis) and see what factors made them similar to each other.

# Part 1

## Load Packages

As the analysis is performed using R environment, certain libraries are used to perform data mining on the data set. So, the first part of the analysis is to bring the predefined package libraries required for our analysis to work environment. A comment is mentioned next to the library about the use of that package in the project.

library(tidyverse) # used for data manupulation and visulization

## ── Attaching packages ────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.0.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.6  
## ✔ tidyr 0.8.1 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(lubridate) # Used for date operations

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(statsr) # Statisticel Inference

## Loading required package: BayesFactor

## Loading required package: coda

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':  
##   
## expand

## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).  
##   
## Type BFManual() to open the manual.  
## \*\*\*\*\*\*\*\*\*\*\*\*

library(reshape2) # Data Shaping and Scaling

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

library(cluster) # Cluster analysis  
library(vcd)# For liklehood ratio in chisquare tes

## Loading required package: grid

# Part 2

## Load Data

Function read.csv() loads the data and shows what data type are the columns. If a column datatype is not as required for the analysis like character() instead of a double(), this is the stage to identify them.

data1 <- read.csv("Propphilly.csv")

# Part 3

## Cleaning the data

It is important to clean the data and represent the fields what they mean by, in this part data cleaning is performed

In the data provided some columns are represented by wrong data types and special characters are present in the data like $ sign, so they needs to be cleaned and represented in a correct data type to apply the appropriate data mining techniques.

Comments are provided against each cleaning/modeling step.

# Deleting $ and , sign from sale price and creating a new column SoldPrice as numeric datatype  
data1$SoldPrice <- as.numeric(gsub("[\\$,]", "", data1$SalePrice))  
is.numeric(data1$SoldPrice)

## [1] TRUE

# Converting Sale date from Charcter to Date   
data1$SaleDate <- as.Date(data1$SaleDate,"%B %d %Y")  
is.Date(data1$SaleDate)

## [1] TRUE

# Converting OpeningBid from Character to Numeric  
data1$OpeningBid <- as.numeric(as.character(data1$OpeningBid))  
is.numeric(data1$OpeningBid)

## [1] TRUE

# Converting Bathrooms from Character to Numeric  
suppressWarnings(data1$bathrooms <- as.numeric(as.character(data1$bathrooms)))  
is.numeric(data1$bathrooms)

## [1] TRUE

# Converting Bedrooms from Character to Numeric  
suppressWarnings(data1$bedrooms <- as.numeric(as.character(data1$bedrooms)))  
is.numeric(data1$bathrooms)

## [1] TRUE

#Converting Postal Code as character vector  
data1$PostalCode <- as.character(as.numeric(data1$PostalCode))  
is.character(data1$PostalCode)

## [1] TRUE

#Converting Ward Number as character vector  
data1$Ward <- as.character(as.numeric(data1$Ward))  
is.character(data1$Ward)

## [1] TRUE

# Part 4

## An overview of the data

Total number of houses in the data (rows)

#Number of rows in the  
nrow(data1)

## [1] 391

Total features about each house (columns)

#Number of columns  
ncol(data1)

## [1] 28

Summary of the data (Descriptive Statistics on all the features in the data)

#Summary of the data (Provides the summary of all the columns like mean, median, etc.)  
summary(data1)

## Address   
## 1033 E HAINES ST : 1   
## 110 W CHAMPLOST AVE : 1   
## 1105 KENWYN ST : 1   
## 1109 DIVINITY ST : 1   
## 1109 OVERINGTON ST : 1   
## 111 W ABBOTTSFORD AVE : 1   
## (Other) :385

## ZillowAddress SaleDate   
## 1033 E HAINES ST , Philadelphia, PA 19138 : 1 Min. :2016-08-02   
## 110 W CHAMPLOST AVE , Philadelphia, PA 19120 : 1 1st Qu.:2016-08-02   
## 1105 KENWYN ST , Philadelphia, PA 19124 : 1 Median :2016-09-13   
## 1109 DIVINITY ST , Philadelphia, PA 19143 : 1 Mean :2016-08-30   
## 1109 OVERINGTON ST , Philadelphia, PA 19124 : 1 3rd Qu.:2016-09-13   
## 111 W ABBOTTSFORD AVE , Philadelphia, PA 19144: 1 Max. :2017-03-07   
## (Other) :385   
## OpeningBid SalePrice OPA PostalCode   
## Min. : 2200 $40,000 : 17 Min. : 11241900 Length:391   
## 1st Qu.: 8950 $20,000 : 15 1st Qu.:251353750 Class :character   
## Median :11300 $30,000 : 13 Median :406166400 Mode :character   
## Mean :13230 $60,000 : 13 Mean :388502828   
## 3rd Qu.:14400 $25,000 : 11 3rd Qu.:522163450   
## Max. :64800 $125,000 : 9 Max. :888651465   
## (Other) :313   
## Attorney Ward   
## PHELAN HALLINAN LLP :102 Length:391   
## KML LAW GROUP : 72 Class :character   
## MANLEY DEAS KOCHALSKI LLC : 41 Mode :character   
## MCCABE WEISBERG & CONWAY P.C.: 28   
## UDREN LAW OFFICES P.C. : 25   
## MILSTEAD & ASSOCIATES LLC : 20   
## (Other) :103   
## Seller   
## WELLS FARGO BANK N.A. : 20   
## WELLS FARGO BANK NA : 18   
## FEDERAL NATIONAL MORTGAGE ASSOCIATION: 12   
## MTGLQ INVESTORS L.P. : 12   
## BANK OF AMERICA N.A. : 9   
## LSF9 MASTER PARTICIPATION TRUST : 9   
## (Other) :311   
## Buyer SheriffCost Advertising   
## PHELAN HALLINAN LLP : 44 Min. :1059 Min. :1455   
## KML LAW GROUP : 29 1st Qu.:1989 1st Qu.:1723   
## FEDERAL NATIONAL MORTGAGE ASSOCIATION: 17 Median :2539 Median :1784   
## MANLEY DEAS KOCHALSKI LLC : 17 Mean :2898 Mean :1758   
## UDREN LAW OFFICES P.C. : 14 3rd Qu.:3500 3rd Qu.:1811   
## MILSTEAD & ASSOCIATES LLC : 13 Max. :9712 Max. :2055   
## (Other) :257   
## Water PGW AverageWalkAndTransitScore  
## Min. : 23.79 Min. : 0.44 Min. :45.00   
## 1st Qu.: 415.73 1st Qu.: 371.14 1st Qu.:66.00   
## Median : 959.27 Median : 1079.18 Median :71.25   
## Mean : 2274.95 Mean : 2127.50 Mean :70.69   
## 3rd Qu.: 2400.66 3rd Qu.: 2537.51 3rd Qu.:79.00   
## Max. :53187.97 Max. :19371.97 Max. :90.00   
##

## ViolentCrimeRate SchoolScore ZillowEstimate RentEstimate   
## Min. :0.0400 Min. : 3.34 Min. : 20806 Min. : 800   
## 1st Qu.:0.3600 1st Qu.: 7.89 1st Qu.: 71298 1st Qu.:1100   
## Median :0.5200 Median :10.94 Median :104652 Median :1200   
## Mean :0.6477 Mean :14.16 Mean :126198 Mean :1239   
## 3rd Qu.:0.8400 3rd Qu.:17.00 3rd Qu.:161024 3rd Qu.:1312   
## Max. :1.7200 Max. :69.50 Max. :781509 Max. :4581   
##   
## taxAssessment yearBuilt finishedSqft bathrooms   
## Min. : 13400 Min. : 0 Min. : 1 Min. :1.000   
## 1st Qu.: 65300 1st Qu.:1923 1st Qu.:1072 1st Qu.:1.000   
## Median : 95200 Median :1925 Median :1209 Median :1.000   
## Mean :108393 Mean :1900 Mean :1320 Mean :1.387   
## 3rd Qu.:138950 3rd Qu.:1950 3rd Qu.:1440 3rd Qu.:2.000   
## Max. :517200 Max. :2006 Max. :4564 Max. :4.000   
##   
## bedrooms PropType Averagecomps SoldPrice   
## Min. :1.00 Condominium :116 Min. : 21167 Min. : 6200   
## 1st Qu.:3.00 MultiFamily2To4: 4 1st Qu.: 72210 1st Qu.: 25000   
## Median :3.00 SingleFamily :124 Median :103283 Median : 45000   
## Mean :3.11 Townhouse :147 Mean :119775 Mean : 62575   
## 3rd Qu.:3.00 3rd Qu.:148890 3rd Qu.: 85000   
## Max. :6.00 Max. :790500 Max. :350000   
##

# Part 5

## Analysis 1

# Multiple Linear Regression

### Motivation

In order to predict the Sold Price of a house in Philadelphia. A Multiple Linear Regression is used. In this analysis we can find which features are significantly affecting the price of a house in Philly.

### Hypothesis

#### Null Hypothesis (H0):

All the Variables does not help to predict SoldPrice in the model

#### Alternative Hypothesis (HA):

At least one variable help to predict SoldPrice in the model

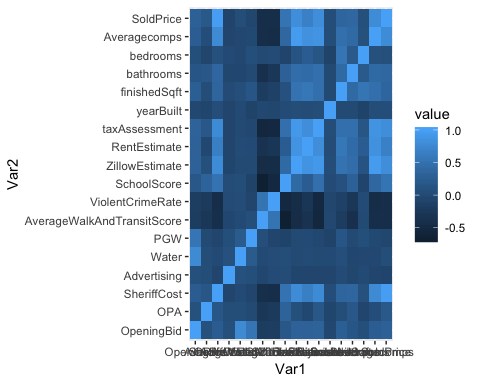
### Pearson Correlation and Heatmap

A correlation heat map shows how high the variables are correlated to each other, this helps to sort the over correlated predictors and check whether a factor analysis is required.

#Selecting the numeric data  
only.numeric <- data1[,sapply(data1, is.numeric)]  
  
#Correlation   
x1cor <- round(cor(only.numeric),2)  
x2cor <- melt(x1cor)  
  
#Creating the heat map  
ggheatmap <- ggplot(data = x2cor, aes(x=Var1, y=Var2, fill=value)) +   
 geom\_tile()  
  
#Visualizing the heatmap is preferred instead of printing the complete correlation matrix (For correlation results please use the R markdown file)

x1cor

ggheatmap



After checking the correlation values and the heat map , SheriffCost is over Correlated (about 100%) with y - soldprice and other x variables hence, when building a linear model it is better to remove SheriffCost predictor for a practical values.

AS observed in the plotsthere is non-variance issue with many predictors. So three models are considered for comparison in the initial phase. (Plots are removed from the document as they occupy more space, for viewing the plots please use R markdown file)

1. Normal model - SoldPrice is predicted
2. Square root model - Square root of SoldPrice is predicted
3. Natural log model - log(SoldPrice) is predicted

Which ever model has the highest R-square value is selected for further step-wise regression

In this step we will predict the Sold Price using the significant predictors. A step wise regression is performed to eliminate the non significant variables after selection of the compared models.

**Step - 1:**

3 models with all the x variables

## modelN – Normal Model  
modelN <- lm(SoldPrice ~ OpeningBid + SaleDate + OPA + Advertising + Water + PGW + AverageWalkAndTransitScore + ViolentCrimeRate + SchoolScore + ZillowEstimate + RentEstimate + taxAssessment + yearBuilt + finishedSqft + bathrooms + bedrooms + Averagecomps , data = data1)  
  
summary(modelN)

## Results  
   
## Residuals:  
## Min 1Q Median 3Q Max   
## -117374 -13200 397 13990 153706   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.723e+05 8.899e+05 -0.531 0.595881   
## OpeningBid 5.600e-01 5.058e-01 1.107 0.268905   
## SaleDate 2.446e+01 5.194e+01 0.471 0.638055   
## OPA 1.526e-05 8.190e-06 1.864 0.063158 .   
## Advertising -1.840e+01 1.308e+01 -1.407 0.160401   
## Water -1.892e-02 6.194e-01 -0.031 0.975648   
## PGW -7.332e-01 7.323e-01 -1.001 0.317386   
## AverageWalkAndTransitScore 7.817e+02 2.331e+02 3.353 0.000880 \*\*\*  
## ViolentCrimeRate 3.778e+03 4.737e+03 0.798 0.425629   
## SchoolScore 1.092e+03 1.776e+02 6.149 2e-09 \*\*\*  
## ZillowEstimate 2.424e-01 7.089e-02 3.419 0.000698 \*\*\*  
## RentEstimate -1.828e+01 1.059e+01 -1.726 0.085135 .   
## taxAssessment 2.174e-01 6.843e-02 3.177 0.001613 \*\*   
## yearBuilt 4.264e+00 5.386e+00 0.792 0.429035   
## finishedSqft -8.017e+00 4.597e+00 -1.744 0.082006 .   
## bathrooms 3.961e+02 2.486e+03 0.159 0.873500   
## bedrooms 4.831e+03 2.618e+03 1.845 0.065781 .   
## Averagecomps 1.750e-01 5.784e-02 3.026 0.002649 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26800 on 373 degrees of freedom  
## Multiple R-squared: 0.7373, Adjusted R-squared: 0.7254   
## F-statistic: 61.6 on 17 and 373 DF, p-value: < 2.2e-16

## modelsqrt – Square Root Model

modelsqrt <- lm(sqrt(SoldPrice) ~ OpeningBid + SaleDate + OPA + Advertising + Water + PGW + AverageWalkAndTransitScore + ViolentCrimeRate + SchoolScore + ZillowEstimate + RentEstimate + taxAssessment + yearBuilt + finishedSqft + bathrooms + bedrooms + Averagecomps , data = data1)  
  
summary(modelsqrt)

## Results  
  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -226.735 -28.072 4.123 33.614 250.152   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.416e+03 1.838e+03 -0.770 0.441561   
## OpeningBid 7.832e-04 1.045e-03 0.750 0.453852   
## SaleDate 8.831e-02 1.073e-01 0.823 0.410924   
## OPA 3.063e-08 1.691e-08 1.811 0.070927 .   
## Advertising -3.018e-02 2.702e-02 -1.117 0.264802   
## Water 4.572e-04 1.279e-03 0.357 0.720967   
## PGW -7.379e-04 1.512e-03 -0.488 0.625912   
## AverageWalkAndTransitScore 8.207e-01 4.814e-01 1.705 0.089104 .   
## ViolentCrimeRate -7.384e+00 9.784e+00 -0.755 0.450858   
## SchoolScore 1.625e+00 3.667e-01 4.431 1.24e-05 \*\*\*  
## ZillowEstimate 3.683e-04 1.464e-04 2.516 0.012296 \*   
## RentEstimate -6.225e-02 2.187e-02 -2.846 0.004674 \*\*   
## taxAssessment 4.048e-04 1.413e-04 2.865 0.004412 \*\*   
## yearBuilt 8.625e-03 1.112e-02 0.775 0.438622   
## finishedSqft -2.240e-02 9.495e-03 -2.359 0.018841 \*   
## bathrooms 2.866e+00 5.134e+00 0.558 0.576940   
## bedrooms 1.304e+01 5.407e+00 2.412 0.016325 \*   
## Averagecomps 4.369e-04 1.194e-04 3.658 0.000291 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 55.34 on 373 degrees of freedom  
## Multiple R-squared: 0.6808, Adjusted R-squared: 0.6662   
## F-statistic: 46.79 on 17 and 373 DF, p-value: < 2.2e-16

## modellog – Natural Log Model

modellog <- lm(log(SoldPrice) ~ OpeningBid + SaleDate + OPA + Advertising + Water + PGW + AverageWalkAndTransitScore + ViolentCrimeRate + SchoolScore + ZillowEstimate + RentEstimate + taxAssessment + yearBuilt + finishedSqft + bathrooms + bedrooms + Averagecomps , data = data1)  
  
summary(modellog)

## Result  
  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0644 -0.2503 0.1033 0.3579 1.7667   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.381e+00 1.923e+01 -0.488 0.626027   
## OpeningBid 6.250e-06 1.093e-05 0.572 0.567838   
## SaleDate 1.137e-03 1.123e-03 1.013 0.311842   
## OPA 2.913e-10 1.770e-10 1.646 0.100674   
## Advertising -2.433e-04 2.828e-04 -0.860 0.390158   
## Water 7.073e-06 1.339e-05 0.528 0.597598   
## PGW -1.121e-06 1.583e-05 -0.071 0.943594   
## AverageWalkAndTransitScore 4.361e-03 5.038e-03 0.866 0.387237   
## ViolentCrimeRate -2.235e-01 1.024e-01 -2.183 0.029665 \*   
## SchoolScore 1.001e-02 3.838e-03 2.607 0.009506 \*\*   
## ZillowEstimate 2.342e-06 1.532e-06 1.529 0.127206   
## RentEstimate -6.822e-04 2.289e-04 -2.980 0.003069 \*\*   
## taxAssessment 3.442e-06 1.479e-06 2.328 0.020469 \*   
## yearBuilt 6.010e-05 1.164e-04 0.516 0.605996   
## finishedSqft -2.392e-04 9.936e-05 -2.407 0.016572 \*   
## bathrooms 3.214e-02 5.373e-02 0.598 0.550113   
## bedrooms 1.373e-01 5.659e-02 2.426 0.015742 \*   
## Averagecomps 4.428e-06 1.250e-06 3.542 0.000447 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5792 on 373 degrees of freedom  
## Multiple R-squared: 0.5692, Adjusted R-squared: 0.5496   
## F-statistic: 28.99 on 17 and 373 DF, p-value: < 2.2e-16

## Among all the variables AverageWalkAndTransitScore , SchoolScore, ZillowEstimate, taxAssessment, and Averagecomps

**Step -2**

The regular model has the highest R square value of 0.7251 i.e 72.51% and the significant x variables(Predictors) are , School Score, ZillowEstimate, taxAssessment, and Averagecomps (P - values of these variables are less than 0.05)

modelN <- lm(SoldPrice ~ AverageWalkAndTransitScore + SchoolScore + ZillowEstimate + taxAssessment + Averagecomps , data = data1)  
  
summary(modelN)

## Results

## Call:  
## lm(formula = SoldPrice ~ AverageWalkAndTransitScore + SchoolScore +   
## ZillowEstimate + taxAssessment + Averagecomps, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -114755 -13064 -542 14134 146744   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.063e+04 1.869e+04 -4.314 2.04e-05 \*\*\*  
## AverageWalkAndTransitScore 8.286e+02 2.271e+02 3.649 0.0003 \*\*\*  
## SchoolScore 1.221e+03 1.665e+02 7.337 1.31e-12 \*\*\*  
## ZillowEstimate 1.845e-01 6.715e-02 2.747 0.0063 \*\*   
## taxAssessment 2.602e-01 6.213e-02 4.188 3.49e-05 \*\*\*  
## Averagecomps 1.323e-01 5.602e-02 2.362 0.0187 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 27000 on 385 degrees of freedom  
## Multiple R-squared: 0.7249, Adjusted R-squared: 0.7213   
## F-statistic: 202.9 on 5 and 385 DF, p-value: < 2.2e-16

#### Result

Null hypothesis is rejected as there are 5 significant variables help to predict SoldPrice in the model.

AverageWalkAndTransitScore , SchoolScore, ZillowEstimate, taxAssessment, and Averagecomps are the good predictors of SoldPrice of a house.

#### Prediction Equation

SoldPrice in $’s (Y) = (-80630) + 828.6(AverageWalkAndTransitScore) + 1221(SchoolScore) + 0.1845(ZillowEstimate) + 0.2602(taxAssessment) + 0.1323(Averagecomps)

#### How does the significant predictors effect SoldPrice of a house -

1. For ever increase of one point of Average Walk and Transit Score, the sold price of the house is increased by 828.6 dollars
2. For ever increase of one point of SchoolScore, the sold price of the house is increased by 1221 dollars
3. For ever increase of one dollar of ZillowEstimate, the sold price of the house is increased by 0.1845 dollars
4. For ever increase of one dollar of Tax Assessment, the sold price of the house is increased by 0.2602 dollars
5. For ever increase of Averagecomps, the sold price of the house is increased by 0.1323 dollars

## 

## Analysis 2

# ANNOVA

### Why Annova -

1 . ANNOVA is the best approach when comparing means of categories

2 . We have the to be predicted Prop Type a categorical data and predict sold price continuous data, Analysis of the variance is appropriate for this type of analysis.

### Motivation -

The reason to perform an Analysis of Variance test is to find whether the type of house (PropType) is effecting the mean sold price of the house.

### Hypothesis -

#### Null Hypothesis (H0):

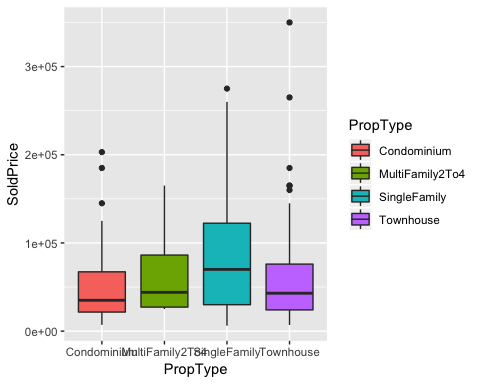
The mean cost of all type of houses (PropType) is same

#### Alternative Hypothesis (HA):

The mean cost of at least one house type (PropType) is different

Variability test

ggplot(data1,aes(x=PropType,y=SoldPrice)) + geom\_boxplot(aes(fill=PropType))



In the above plot the variability across the groups looks about to be equal

### Annova test

anova(lm(SoldPrice ~ PropType, data= data1))

## Analysis of Variance Table  
##   
## Response: SoldPrice  
## Df Sum Sq Mean Sq F value Pr(>F)   
## PropType 3 7.2361e+10 2.4120e+10 9.8526 2.825e-06 \*\*\*  
## Residuals 387 9.4742e+11 2.4481e+09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Result - As the P-Value is very less than level of significance 0.05 hence the null hypothesis is rejected.

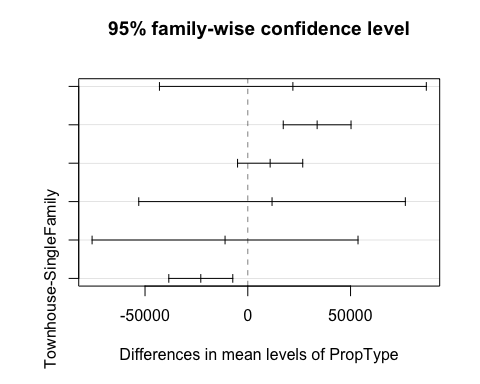
We are 95% confident that the mean cost of at least one house type (PropType) is different.

### Tukey HSD test –

x12 <- aov(lm(SoldPrice ~ PropType, data= data1))  
y12 <- TukeyHSD(x12, conf.level = 0.95)  
y12

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = lm(SoldPrice ~ PropType, data = data1))  
##   
## $PropType  
## diff lwr upr p adj  
## MultiFamily2To4-Condominium 21941.38 -42982.600 86865.359 0.8193445  
## SingleFamily-Condominium 33745.41 17254.704 50236.119 0.0000013  
## Townhouse-Condominium 10879.47 -4975.433 26734.382 0.2891639  
## SingleFamily-MultiFamily2To4 11804.03 -53050.099 76658.163 0.9656810  
## Townhouse-MultiFamily2To4 -11061.90 -75757.291 53633.481 0.9712721  
## Townhouse-SingleFamily -22865.94 -38432.353 -7299.521 0.0009993

plot(y12)



From the Tu-key’s HSD,

A. House types(PropType) SingleFamily and Condominium have a significance difference between their mean SoldPrice. We are 95% confident that SingleFamily sold price is higher than Condominium between $17255 to $50236

B. House types(PropType) SingleFamily and Townhouse have a significance difference between their mean SoldPrice. We are 95% confident that SingleFamily sold price is higher than Townhouse between $7300 to $38432

## Analysis 3

# Chi-Square test of independence

#### Why Chi-Square

As we are comparing two categorical variables Chi-Square test of independence is an ideal way of approach.

#### Hypothesis

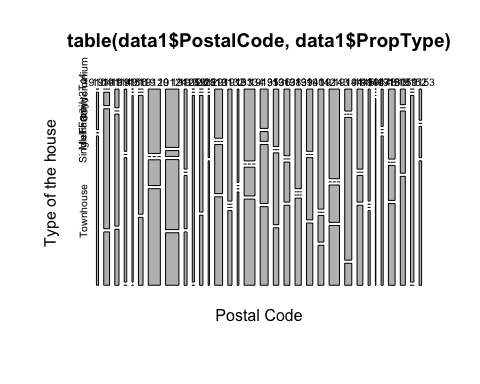
Null Hypothesis (H0) - The type of the house (PropType) is independent of the location (Postal Code) of the house

Alternative Hypothesis (HA) - The type of the house (PropType) is dependent of the location (Postal Code) of the house

### Mosaic Plot

A mosaic plot shows the comparison between the two categorical variables

plot(table(data1$PostalCode,data1$PropType), xlab="Postal Code", ylab="Type of the house")



The plot looks cluttered due to many postal codes in the data, but when we compare the categories in the plot. There is difference in the mosaics between the postal codes (The mosaic tiles tell the number of specific type of house in that postal code)

chisq.test(data1$PropType, data1$PostalCode) ## Pearson test

## Warning in chisq.test(data1$PropType, data1$PostalCode): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: data1$PropType and data1$PostalCode  
## X-squared = 175.42, df = 93, p-value = 5.2e-07

# Likelyhood ratio  
xz123 <- xtabs(~PropType + PostalCode, data = data1)  
assocstats(xz123)

## X^2 df P(> X^2)  
## Likelihood Ratio 184.12 93 5.8086e-08  
## Pearson 175.42 93 5.2001e-07  
##

## Phi-Coefficient : NA   
## Contingency Coeff.: 0.557   
## Cramer's V : 0.387

Result - The Likelihood Ratio is smaller than the level of significance, thus the null hypothesis is rejected. Thus the type of the property is dependent on the location of the property.

## Analysis 4

# Cluster Analysis

Cluster analysis is used to find which houses in the data are similar.

#### Step - 1

Cluster analysis need only numeric data so select the numeric data from the data set

clus1 <- select\_if(data1, is.numeric)  
summary(clus1)

## OpeningBid OPA SheriffCost Advertising   
## Min. : 2200 Min. : 11241900 Min. :1059 Min. :1455   
## 1st Qu.: 8950 1st Qu.:251353750 1st Qu.:1989 1st Qu.:1723   
## Median :11300 Median :406166400 Median :2539 Median :1784   
## Mean :13230 Mean :388502828 Mean :2898 Mean :1758   
## 3rd Qu.:14400 3rd Qu.:522163450 3rd Qu.:3500 3rd Qu.:1811   
## Max. :64800 Max. :888651465 Max. :9712 Max. :2055   
## Water PGW AverageWalkAndTransitScore  
## Min. : 23.79 Min. : 0.44 Min. :45.00   
## 1st Qu.: 415.73 1st Qu.: 371.14 1st Qu.:66.00   
## Median : 959.27 Median : 1079.18 Median :71.25   
## Mean : 2274.95 Mean : 2127.50 Mean :70.69   
## 3rd Qu.: 2400.66 3rd Qu.: 2537.51 3rd Qu.:79.00   
## Max. :53187.97 Max. :19371.97 Max. :90.00   
## ViolentCrimeRate SchoolScore ZillowEstimate RentEstimate   
## Min. :0.0400 Min. : 3.34 Min. : 20806 Min. : 800   
## 1st Qu.:0.3600 1st Qu.: 7.89 1st Qu.: 71298 1st Qu.:1100   
## Median :0.5200 Median :10.94 Median :104652 Median :1200   
## Mean :0.6477 Mean :14.16 Mean :126198 Mean :1239   
## 3rd Qu.:0.8400 3rd Qu.:17.00 3rd Qu.:161024 3rd Qu.:1312   
## Max. :1.7200 Max. :69.50 Max. :781509 Max. :4581   
## taxAssessment yearBuilt finishedSqft bathrooms   
## Min. : 13400 Min. : 0 Min. : 1 Min. :1.000   
## 1st Qu.: 65300 1st Qu.:1923 1st Qu.:1072 1st Qu.:1.000   
## Median : 95200 Median :1925 Median :1209 Median :1.000   
## Mean :108393 Mean :1900 Mean :1320 Mean :1.387   
## 3rd Qu.:138950 3rd Qu.:1950 3rd Qu.:1440 3rd Qu.:2.000   
## Max. :517200 Max. :2006 Max. :4564 Max. :4.000   
## bedrooms Averagecomps SoldPrice   
## Min. :1.00 Min. : 21167 Min. : 6200   
## 1st Qu.:3.00 1st Qu.: 72210 1st Qu.: 25000   
## Median :3.00 Median :103283 Median : 45000   
## Mean :3.11 Mean :119775 Mean : 62575   
## 3rd Qu.:3.00 3rd Qu.:148890 3rd Qu.: 85000   
## Max. :6.00 Max. :790500 Max. :350000

#### Step - 2

Standardizing Data - It is imporatant that all the data is stadardized and no feature is getting an extra importance in the analysis

means <- apply(clus1,2,mean)  
sds <- apply(clus1,2,sd)  
clus1 <- scale(clus1,means,sds)

#### Step - 3

Calculating Euclidean Distance - used to plot the Dendogram

distance <- dist(clus1)

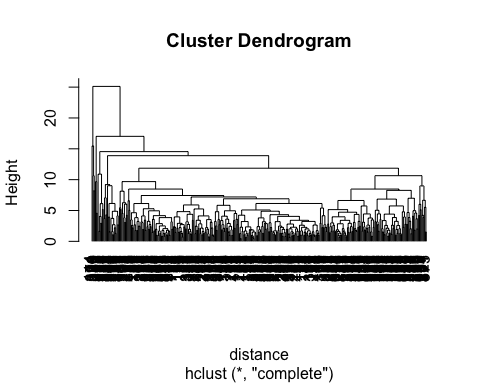
#### Step - 4

Hierarchical Cluster Dendrogram - In the dendograms below we have two types of hierarchical clustering performed in two differnt methods Complete linkage and Average linkage

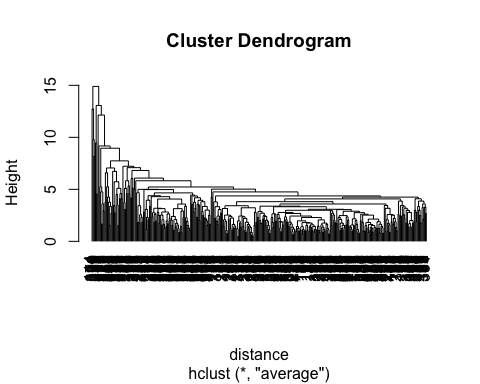
The base(Bottom most) of the dendrogram every individual house is in its own cluster (391 Clusters) as it reaches the top of the plot, houses are merged into single cluster.

The distance between the houses is directly propotional to the dis-similarity between the houses - The more the distance is the less similarity between the houses.

#Complete Linkage  
clus1.com <- hclust(distance,method="complete")  
plot(clus1.com,hang=-1)



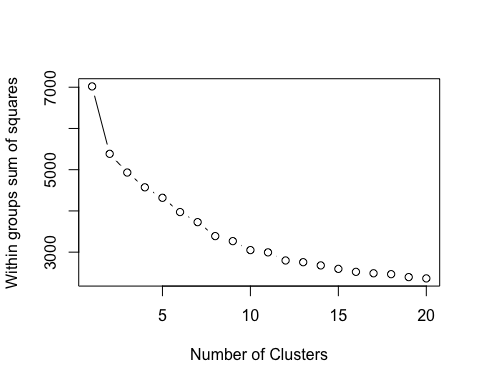
#AVERAGE LINKAGE  
clus1.a <- hclust(distance,method = "average")  
plot(clus1.a,hang=-1)



#### Step - 4

Determining ideal number of clusters using scree plot

# Determine number of clusters  
wss <- (nrow(clus1)-1)\*sum(apply(clus1,2,var))  
for (i in 2:20) wss[i] <- sum(kmeans(clus1,   
 centers=i)$withinss)  
plot(1:20, wss, type="b", xlab="Number of Clusters",  
 ylab="Within groups sum of squares")



The Elbow in the scree plot has raise at 3rd cluster so I chose a 3 cluster solution

#### Step - 5

K means cluster with 3 cluster solution

kc<-kmeans(clus1,3)  
cluscol <- c("blue","green","red") # Used to colour the clusters  
kc

## K-means clustering with 3 clusters of sizes 38, 144, 209  
##   
## Cluster means:  
## OpeningBid OPA SheriffCost Advertising Water PGW  
## 1 0.872841104 0.3701935 1.8668418 -0.07343189 0.19060555 -0.120061314  
## 2 -0.003704504 0.2254175 0.4226030 -0.01768693 -0.13886230 0.027415316  
## 3 -0.156145997 -0.2226195 -0.6305972 0.02553746 0.06101991 0.002940309  
## AverageWalkAndTransitScore ViolentCrimeRate SchoolScore ZillowEstimate  
## 1 -1.5232441 -0.9118722 1.6405196 2.0128367  
## 2 -0.3650170 -0.5379555 0.2217091 0.3348315  
## 3 0.5284485 0.5364437 -0.4510328 -0.5966676  
## RentEstimate taxAssessment yearBuilt finishedSqft bathrooms  
## 1 1.7889236 2.0234457 0.003438694 1.263707721 1.2626736  
## 2 0.1724432 0.3749379 0.152334154 0.009847122 0.2275946  
## 3 -0.4440714 -0.6262296 -0.105582721 -0.236549660 -0.3863886  
## bedrooms Averagecomps SoldPrice  
## 1 0.35767169 1.8620916 1.9653448  
## 2 0.02315170 0.3831321 0.3913192  
## 3 -0.08098262 -0.6025383 -0.6269525  
##   
## Clustering vector:  
## [1] 3 2 3 2 2 2 3 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3  
## [36] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 2  
## [71] 3 3 3 3 2 3 3 3 3 3 3 2 2 3 3 3 2 3 3 3 1 2 3 3 3 3 3 1 3 3 3 3 3 2 3  
## [106] 2 3 3 2 3 3 3 3 1 2 3 2 3 3 2 3 3 3 3 3 3 3 3 3 2 3 2 2 3 3 1 3 2 3 3  
## [141] 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 2 3 2 3 3 2 3 2 2 3 2 3 2 2 2 3 2 3 2  
## [176] 3 3 3 3 3 2 2 2 3 3 3 2 3 2 3 2 3 2 2 3 2 2 3 1 2 3 2 2 2 2 3 3 3 2 2  
## [211] 2 3 3 2 2 2 3 3 2 3 1 2 2 1 2 2 2 3 3 2 3 3 3 3 3 2 1 2 2 2 1 3 2 1 2  
## [246] 3 2 1 2 3 2 2 2 1 2 2 1 3 2 1 3 2 1 3 2 2 2 1 1 3 2 2 2 3 1 2 2 1 3 1  
## [281] 3 3 2 3 2 2 3 3 2 2 3 3 2 2 2 1 2 1 1 2 2 2 2 2 3 2 3 2 1 3 2 2 2 2 2  
## [316] 3 1 2 2 3 1 3 2 3 2 2 1 3 2 2 3 3 2 3 2 2 1 3 1 1 1 3 2 1 2 2 2 2 3 3  
## [351] 3 2 1 2 2 3 1 2 2 3 2 2 3 3 2 3 2 3 2 2 3 1 2 1 1 2 1 2 2 2 3 2 3 3 2  
## [386] 3 3 3 3 3 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 1269.740 1421.540 2240.765  
## (between\_SS / total\_SS = 29.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

#### Brief Summary –

Clustering Vector shows which house falls in which cluster.

Cluster 1 - Blue color - 38 houses fall in this cluster: The mean Sold Price of the houses is the highest in this cluster amongst all other cluster – Expensive Houses

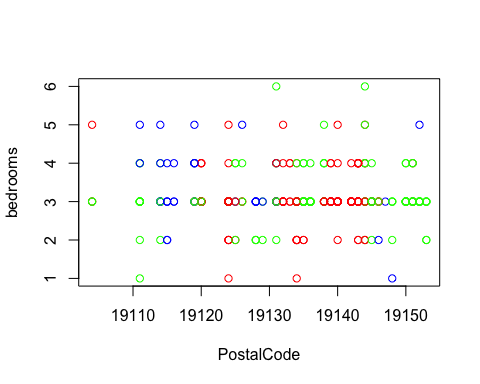
Cluster 2 - Green color - 144 houses fall in this cluster: The mean Sold Price of the houses is less than house in cluster 1 but more than cluster 3 – Medium Priced Houses

Cluster 3 - Red colour - 209 houses fall in this cluster: The mean Sold Price of the houses is less than cluster 2 and cluster 1 – Cheaper Houses

#### Sample Applications

1. Plotting the clusters to postal code and bedrooms

plot(bedrooms~PostalCode, data1, col = cluscol[kc$cluster])



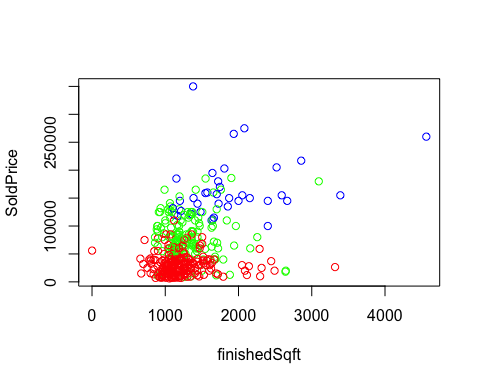
Observation –

Weather the number of bedrooms of a house is more or less, most of the expensive houses (Cluster 1 blue) are in Postal Codes 19110 to 19120

So it is more likely that the price of house in postal codes start 19110 to 19120 are expensive than house prices in 19121 to 19150 in Philadelphia.

1. Plotting the clusters to postal code and bedrooms

plot(SoldPrice~finishedSqft, data1, col = cluscol[kc$cluster])



Observation -

This plot supports the breif summary provided -

House that are smaller and low-priced fall under cluster 3 (Red)

House that are medium sized and medium-priced fall under cluster 2 (Green)

House that are bigger and high-priced fall under cluster 1 (Blue)

# Result

The results for all the analysis performed have been listed at the end of the respective data mining technique performed. Machine learning algorithms like SVM or Random Forrest can be applied for more stable predictions.

\*\*\*\* End of the report \*\*\*\*

Thank you, Dr. Clements, for your valuable lectures