Philadelphia County Real Estate Data 2016-2017

Dong Junduo

August 4, 2017

This data includes 615 records of real estate which were sold by auction in Philadelphia County on Aug 3, Sept 4, and Oct 4 in 2016, and Feb 7 and Mar 7 in 2017, including details of the tranaction in auction, properties details of the real estate, and estimated values from Zillow, etc. Our goal for the following analytics is to look back to the recent sale by acution in the real estate industry and to find the relationships between prices, time, and location, etc. The analytics should also bring insights about the real estate industry in Philadelphia County for business and privates future consideration.

Loading data.

```
setwd("~/Desktop/桌面/Self_done_project/Philadelphia_RealEstate")
Real_estate <-
  read.csv('Properties_philly_Kraggle_v2.csv',
  header = TRUE,
  sep = ",")
```

Overview of data.

```
str(Real_estate)
```

```
## 'data.frame':    615 obs. of  30 variables:
##  $ Address                : Factor w/ 615 levels "1033 E HAINES ST ",..: 162 339 35 7 49 13 232 332 274 53
5 ...
##  $ Zillow.Address         : Factor w/ 615 levels "1033 E HAINES ST , Philadelphia, PA 19138",..: 162 339 3
5 7 49 13 232 332 274 535 ...
##  $ Sale.Date              : Factor w/ 5 levels "August 2  2016",..: 5 1 1 1 1 5 5 4 5 4 ...
##  $ Opening.Bid            : int  11400 8500 12600 9200 8900 8800 10900 10000 10800 7100 ...
##  $ Sale.Price.bid.price   : Factor w/ 229 levels "$10,000 ","$10,100 ",..: 17 208 29 220 211 210 9 1 8 184
...
##  $ Book.Writ              : Factor w/ 615 levels "1501-701","1501-719",..: 406 111 82 168 92 25 469 525 50
0 586 ...
##  $ OPA                    : int  314095100 652092800 592058735 123016300 592159100 531129700 212002100 52
1194100 642138100 406204700 ...
##  $ Postal.Code            : int  19125 19136 19144 19144 19144 19111 19128 19131 19136 19142 ...
##  $ Attorney               : Factor w/ 41 levels "AXEL A. SHIELD  III ",..: 25 24 25 15 25 25 15 25 4 22 ..
.
##  $ Ward                   : int  31 65 59 12 59 53 21 52 64 40 ...
##  $ Seller                 : Factor w/ 313 levels "0","AMERICAN FINANCIAL RESOURCES  INC.",..: 290 206 112
118 290 112 125 290 32 191 ...
##  $ Buyer                  : Factor w/ 274 levels "1168 E HORTTER ST. LLC",..: 178 227 178 114 178 178 114
178 39 151 ...
##  $ Sheriff.Cost           : num  1314 1254 1337 1460 1263 ...
##  $ Advertising            : num  1723 1557 1811 1803 1811 ...
##  $ Other                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Record.Deed            : int  0 0 0 257 0 0 0 0 0 257 ...
##  $ Water                  : num  361 324 452 625 700 ...
##  $ PGW                    : num  415 0 758 188 0 ...
##  $ Avg.Walk.Transit.score: num  82.2 65.2 71.8 71.8 71.8 ...
##  $ Violent.Crime.Rate     : num  1.13 0.35 0.86 0.86 0.86 0.25 0.12 0.29 0.35 0.84 ...
##  $ School.Score           : num  15.87 32.53 7.89 7.89 7.89 ...
##  $ Zillow.Estimate        : Factor w/ 612 levels "100,128.00","100,356.00",..: 253 109 209 178 93 90 220 1
46 111 603 ...
##  $ Rent.Estimate          : Factor w/ 65 levels "1,000.00","1,025.00",..: 33 16 33 28 24 16 24 33 13 6 ...
##  $ taxAssessment          : Factor w/ 511 levels "10,800.00","100,000.00",..: 179 68 203 62 46 70 344 52 1
08 412 ...
##  $ yearBuilt              : int  1921 1959 2006 1900 1925 1947 1955 1925 1963 1925 ...
##  $ finished...SqFt.       : int  898 1120 1882 1792 1419 1260 896 1354 1616 992 ...
##  $ bathrooms              : Factor w/ 9 levels " -   ","1","1.5",..: 2 3 6 2 4 3 2 2 2 2 ...
##  $ bedrooms               : Factor w/ 7 levels " -   ","1","2",..: 3 4 4 5 4 4 3 4 5 4 ...
##  $ PropType               : Factor w/ 4 levels "Condominium",..: 3 4 3 3 3 4 4 3 4 1 ...
##  $ Average.comps          : Factor w/ 599 levels "100,155.56","100,581.78",..: 253 145 241 110 129 86 232
136 94 588 ...
```

Data Cleaning

Looking for missing value(s)

```
data.frame(colSums(is.na(Real_estate)))
```

```
##                         colSums.is.na.Real_estate..
## Address                                          0
## Zillow.Address                                   0
## Sale.Date                                        0
## Opening.Bid                                      5
## Sale.Price.bid.price                             0
## Book.Writ                                        0
## OPA                                              0
## Postal.Code                                      0
## Attorney                                         0
## Ward                                             0
## Seller                                           0
## Buyer                                            0
## Sheriff.Cost                                     0
## Advertising                                      0
## Other                                            0
## Record.Deed                                      0
## Water                                            0
## PGW                                              0
## Avg.Walk.Transit.score                           0
## Violent.Crime.Rate                               0
## School.Score                                     0
## Zillow.Estimate                                  0
## Rent.Estimate                                    0
## taxAssessment                                    0
## yearBuilt                                        0
## finished...SqFt.                                 0
## bathrooms                                        0
## bedrooms                                         0
## PropType                                         0
## Average.comps                                    0
```

From the result shown above, we see there are 5 missing values (NA) in Opening.Bid field.

Replacing the NA value with average value for opening.bid

```
Real_estate$Opening.Bid[which(is.na(Real_estate$Opening.Bid))] <-
  mean(Real_estate$Opening.Bid, na.rm = TRUE)
```

Making sure of no NA value left

```
TRUE %in% is.na(Real_estate$Opening.Bid)
```

```
## [1] FALSE
```

When you look back to the data overview, you see that there are several fields which supposed to be numerical, but storing values as factor, such as Sale.Price.bid.price. A data type of factor would not operate correctly when it comes to calculations. Therefore, we need to convert all such fields to the expected data types.

Converting bid price to numerical format

```
class(Real_estate$Sale.Price.bid.price)
```

```
## [1] "factor"
```

```
Real_estate$Sale.Price.bid.price <-
as.numeric(gsub('[$,]', '', Real_estate$Sale.Price.bid.price))

class(Real_estate$Sale.Price.bid.price)
```

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

Please note that the Sale.Date field is also under factor format, which would be problematic if we need to sort them by dates.

Converting date of sales to date-format

```r
library(stringr)
library(dplyr)

# spliting the field of Sale.Date into three columns separated by space
Sale.Date <-
as.data.frame(str_split_fixed(Real_estate$Sale.Date, ' ', 3))
head(Sale.Date)
```

```
##          V1 V2    V3
## 1 September 13  2016
## 2    August  2  2016
## 3    August  2  2016
## 4    August  2  2016
## 5    August  2  2016
## 6 September 13  2016
```

Changing the data type to characters

```r
Sale.Date[, 1] <- as.character(Sale.Date[, 1]) # month
Sale.Date[, 2] <- as.character(Sale.Date[, 2]) # day
Sale.Date[, 3] <- as.character(Sale.Date[, 3]) # year

# changing the month-field from text to number
Sale.Date[, 1][which(Sale.Date[, 1] == "January")]   <- "01"
Sale.Date[, 1][which(Sale.Date[, 1] == "February")]  <- "02"
Sale.Date[, 1][which(Sale.Date[, 1] == "March")]     <- "03"
Sale.Date[, 1][which(Sale.Date[, 1] == "April")]     <- "04"
Sale.Date[, 1][which(Sale.Date[, 1] == "May")]       <- "05"
Sale.Date[, 1][which(Sale.Date[, 1] == "June")]      <- "06"
Sale.Date[, 1][which(Sale.Date[, 1] == "July")]      <- "07"
Sale.Date[, 1][which(Sale.Date[, 1] == "August")]    <- "08"
Sale.Date[, 1][which(Sale.Date[, 1] == "September")] <- "09"
Sale.Date[, 1][which(Sale.Date[, 1] == "October")]   <- "10"
Sale.Date[, 1][which(Sale.Date[, 1] == "November")]  <- "11"
Sale.Date[, 1][which(Sale.Date[, 1] == "December")]  <- "12"

# naming the columns and add them to the original table Real_estate
Real_estate['Sale.Date_month'] <- Sale.Date[, 1]
Real_estate['Sale.Date_day']   <- Sale.Date[, 2]
Real_estate['Sale.Date_year']  <- Sale.Date[, 3]

# combining the three columns and chaning the data type to date
Real_estate$Sale.Date_format <-
with(Real_estate,
paste0(Sale.Date_year, Sale.Date_month, Sale.Date_day))
Real_estate$Sale.Date_format <-
as.Date(as.character(Real_estate$Sale.Date_format), "%Y%m%d")

# making sure that the new fields are added
str(Real_estate)
```

```
## 'data.frame':    615 obs. of  34 variables:
##  $ Address               : Factor w/ 615 levels "1033 E HAINES ST ",..: 162 339 35 7 49 13 232 332 274 53
5 ...
##  $ Zillow.Address        : Factor w/ 615 levels "1033 E HAINES ST , Philadelphia, PA 19138",..: 162 339 3
5 7 49 13 232 332 274 535 ...
##  $ Sale.Date             : Factor w/ 5 levels "August 2  2016",..: 5 1 1 1 1 5 5 4 5 4 ...
##  $ Opening.Bid           : num  11400 8500 12600 9200 8900 8800 10900 10000 10800 7100 ...
##  $ Sale.Price.bid.price  : num  11400 8500 12600 9200 8900 8800 10900 10000 10800 7100 ...
##  $ Book.Writ             : Factor w/ 615 levels "1501-701","1501-719",..: 406 111 82 168 92 25 469 525 50
0 586 ...
##  $ OPA                   : int  314095100 652092800 592058735 123016300 592159100 531129700 212002100 52
1194100 642138100 406204700 ...
##  $ Postal.Code           : int  19125 19136 19144 19144 19144 19111 19128 19131 19136 19142 ...
##  $ Attorney              : Factor w/ 41 levels "AXEL A. SHIELD  III ",..: 25 24 25 15 25 25 15 25 4 22 ..
.
##  $ Ward                  : int  31 65 59 12 59 53 21 52 64 40 ...
##  $ Seller                : Factor w/ 313 levels "0","AMERICAN FINANCIAL RESOURCES  INC.",..: 290 206 112
118 290 112 125 290 32 191 ...
##  $ Buyer                 : Factor w/ 274 levels "1168 E HORTTER ST. LLC",..: 178 227 178 114 178 178 114
178 39 151 ...
##  $ Sheriff.Cost          : num  1314 1254 1337 1460 1263 ...
##  $ Advertising           : num  1723 1557 1811 1803 1811 ...
##  $ Other                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Record.Deed           : int  0 0 0 257 0 0 0 0 0 257 ...
##  $ Water                 : num  361 324 452 625 700 ...
##  $ PGW                   : num  415 0 758 188 0 ...
##  $ Avg.Walk.Transit.score: num  82.2 65.2 71.8 71.8 71.8 ...
##  $ Violent.Crime.Rate    : num  1.13 0.35 0.86 0.86 0.86 0.25 0.12 0.29 0.35 0.84 ...
##  $ School.Score          : num  15.87 32.53 7.89 7.89 7.89 ...
##  $ Zillow.Estimate       : Factor w/ 612 levels "100,128.00","100,356.00",..: 253 109 209 178 93 90 220 1
46 111 603 ...
##  $ Rent.Estimate         : Factor w/ 65 levels "1,000.00","1,025.00",..: 33 16 33 28 24 16 24 33 13 6 ...
##  $ taxAssessment         : Factor w/ 511 levels "10,800.00","100,000.00",..: 179 68 203 62 46 70 344 52 1
08 412 ...
##  $ yearBuilt             : int  1921 1959 2006 1900 1925 1947 1955 1925 1963 1925 ...
##  $ finished...SqFt.      : int  898 1120 1882 1792 1419 1260 896 1354 1616 992 ...
##  $ bathrooms             : Factor w/ 9 levels " -   ","1","1.5",..: 2 3 6 2 4 3 2 2 2 2 ...
##  $ bedrooms              : Factor w/ 7 levels " -   ","1","2",..: 3 4 4 5 4 4 3 4 5 4 ...
##  $ PropType              : Factor w/ 4 levels "Condominium",..: 3 4 3 3 3 4 4 3 4 1 ...
##  $ Average.comps         : Factor w/ 599 levels "100,155.56","100,581.78",..: 253 145 241 110 129 86 232
136 94 588 ...
##  $ Sale.Date_month       : chr  "09" "08" "08" "08" ...
##  $ Sale.Date_day         : chr  "13" "2" "2" "2" ...
##  $ Sale.Date_year        : chr  " 2016" " 2016" " 2016" " 2016" ...
##  $ Sale.Date_format      : Date, format: "2016-09-13" "2016-08-02" ...
```

Converting more fields to numeric format

```
# converting zillow.Estimate
Real_estate$Zillow.Estimate <- as.numeric(gsub(",", "", Real_estate$Zillow.Estimate))
class(Real_estate$Zillow.Estimate)
```

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

```
# converting rent.estimate
Real_estate$Rent.Estimate   <- as.numeric(gsub(",", "", Real_estate$Rent.Estimate))
class(Real_estate$Rent.Estimate)
```

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

```
# converting taxassessment
Real_estate$taxAssessment   <- as.numeric(gsub(",", "", Real_estate$taxAssessment))
class(Real_estate$taxAssessment)
```

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

```
# converting bathrooms
Real_estate$bathrooms       <- as.numeric(as.character(sub(",",".", Real_estate$bathrooms)))
class(Real_estate$bathrooms)
```

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

```
# converting bedrooms
Real_estate$bedrooms        <- as.numeric(as.character(sub(",",".", Real_estate$bedrooms)))
class(Real_estate$bedrooms)
```

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

```
# converting average.comps
Real_estate$Average.comps   <- as.numeric(gsub(",","", Real_estate$Average.comps))
class(Real_estate$Average.comps)
```

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

Here we need to check again for missing values, since it is possible to find NA after conversion to numeric date types.

```
data.frame(colSums(is.na(Real_estate)))
```

```
##                         colSums.is.na.Real_estate..
## Address                                           0
## Zillow.Address                                    0
## Sale.Date                                         0
## Opening.Bid                                       0
## Sale.Price.bid.price                              0
## Book.Writ                                         0
## OPA                                               0
## Postal.Code                                       0
## Attorney                                          0
## Ward                                              0
## Seller                                            0
## Buyer                                             0
## Sheriff.Cost                                      0
## Advertising                                       0
## Other                                             0
## Record.Deed                                       0
## Water                                             0
## PGW                                               0
## Avg.Walk.Transit.score                            0
## Violent.Crime.Rate                                0
## School.Score                                      0
## Zillow.Estimate                                   0
## Rent.Estimate                                     0
## taxAssessment                                     0
## yearBuilt                                         0
## finished...SqFt.                                  0
## bathrooms                                        35
## bedrooms                                         30
## PropType                                          0
## Average.comps                                     0
## Sale.Date_month                                   0
## Sale.Date_day                                     0
## Sale.Date_year                                    0
## Sale.Date_format                                  0
```

we saw that bathrooms-field and bedrooms-field are now holding missing values. This happens because there were “-” values before converision of data type, and it represents missing values as string. After conversion, the “-” values automatically became “NA.”

Converting NA value for bathrooms

```
Real_estate$bathrooms[which(is.na(Real_estate$bathrooms))] <- mean(Real_estate$bathrooms, na.rm = TRUE)

Real_estate$bathrooms <- format(round(Real_estate$bathrooms, 2),
                                nsmall = 2)

TRUE %in% is.na(Real_estate$bathrooms)
```

```
## [1] FALSE
```

Converting NA value for bedrooms

```
Real_estate$bedrooms[which(is.na(Real_estate$bedrooms))] <-
  mean(Real_estate$bedrooms,
       na.rm = TRUE)

Real_estate$bedrooms <-
  format(round(Real_estate$bedrooms, 2),
         nsmall = 2)

TRUE %in% is.na(Real_estate$bedrooms)
```

```
## [1] FALSE
```

Since it is mentioned that one of the goal for this analytics is to explore how location affect the real estate pricing in Philadelphia County, the next task is to locate the zipcodes in the data. According to the information from http://wikitravel.org/en/Talk:Philadelphia, the analyst decided to split the county into 6 districts: Upper North, Northeast, North, Center, South, and West. So instead of looking at numerous zip codes, the analytics may be produced based on more explicit and descriptive district names.

Creating new distribution of districts for postal codes

```
dataPostalCodes  <- as.numeric(levels(factor(Real_estate$Postal.Code)))
stats_PostalCode <- tapply(Real_estate$Postal.Code, Real_estate$Postal.Code, FUN = length)
stats_PostalCode <- data.frame(stats_PostalCode)

upperNorth <- c(19118, 19119, 19120, 19126, 19127, 19128,
                19129, 19138, 19140, 19141, 19144, 19150)
northEast  <- c(19111, 19114, 19115, 19116, 19124, 19125,
                19134, 19135, 19136, 19137, 19149, 19152, 19154)
north      <- c(19121, 19122, 19123, 19130, 19132, 19133)
center     <- c(19102, 19103, 19106, 19107)
south      <- c(19112, 19145, 19146, 19147, 19148)
west       <- c(19104, 19131, 19139, 19142, 19143, 19151, 19153)

realPostalCodes <- sort(c(upperNorth,
                          northEast,
                          north,
                          center,
                          south,
                          west))

# checking if all postal codes from the data are included in the actual Philly zipmap
all(dataPostalCodes %in% realPostalCodes)
```

```
## [1] TRUE
```

Determining which district is each record located

```r
checkDistrict <- function(zipCode){
  if(zipCode %in% upperNorth){
    return("Upper North")
  }
  else if(zipCode %in% northEast){
    return("Northeast")
  }
  else if(zipCode %in% north){
    return("North")
  }
  else if(zipCode %in% center){
    return("Center")
  }
  else if(zipCode %in% south){
    return("South")
  }
  else if(zipCode %in% west){
    return("West")
  }
  else{
    return("ERROR: THIS IS NOT IN PHILADELPHIA COUNTY!")
  }
}

for(i in 1:length(Real_estate$Postal.Code)){
  Real_estate$District[i] <- checkDistrict(Real_estate$Postal.Code[i])
}

levels(factor(Real_estate$District))
```

```
## [1] "North"       "Northeast"   "South"       "Upper North" "West"
```

Now it is clear that there was no sales by auction of real estate in the data from Center Philly Before the data cleaning comes to its conclusion, we need to make sure again that there is no missing value in the data.

```r
TRUE %in% is.na(Real_estate)
```

```
## [1] FALSE
```

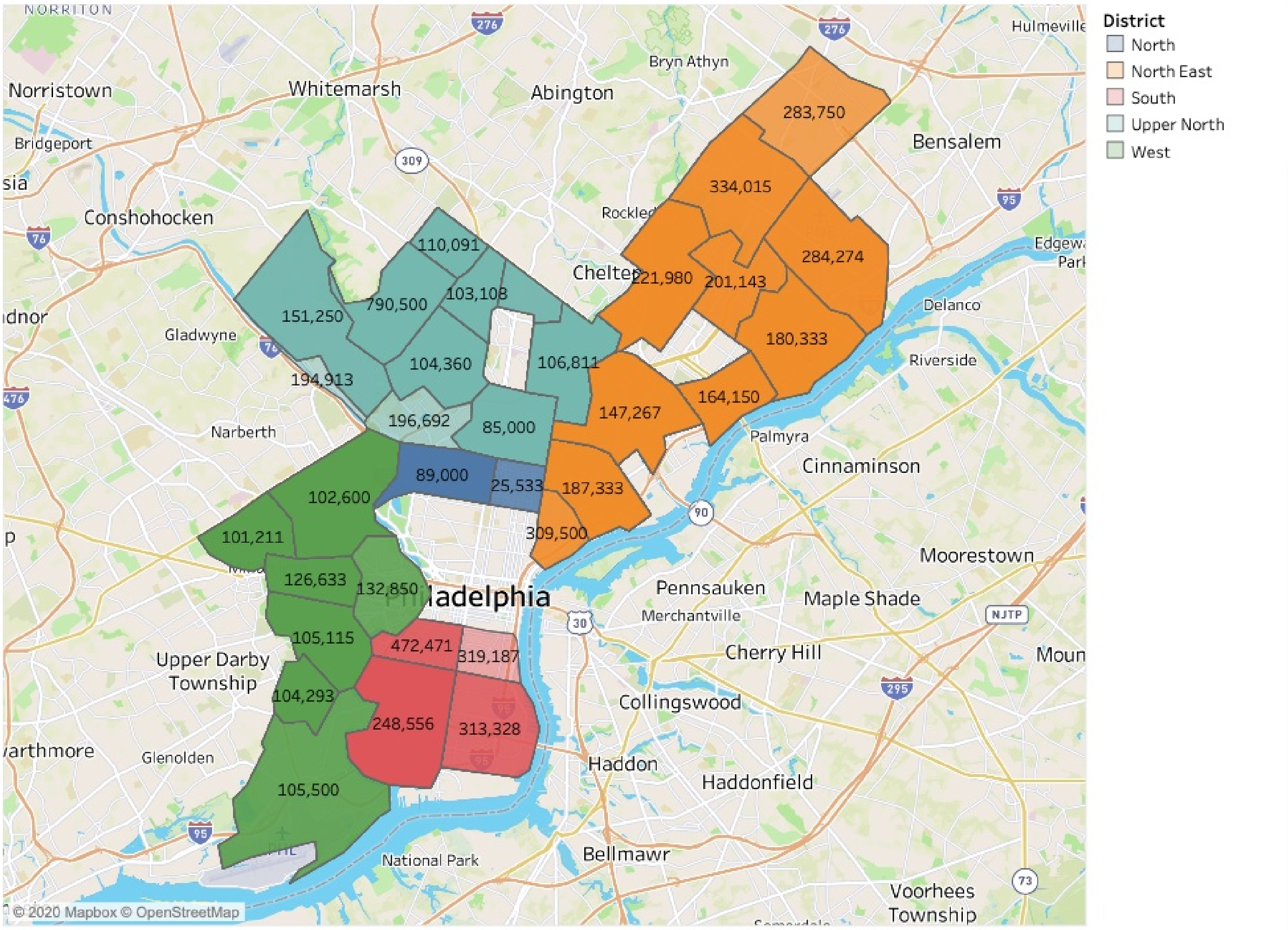
Explotory Data Analysis

Beginning with ploting the locations and the districts in a Google Map. Before ploting the map, langitude and latitude are needed. By using ggmap-function, a package in R, we would be able to acquire the desired information.

(D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URLhttp://journal.r-project.org/archive/2013-1/kahle-wickham.pdf)

```r
library(ggplot2)
```

Philadelphia County Map

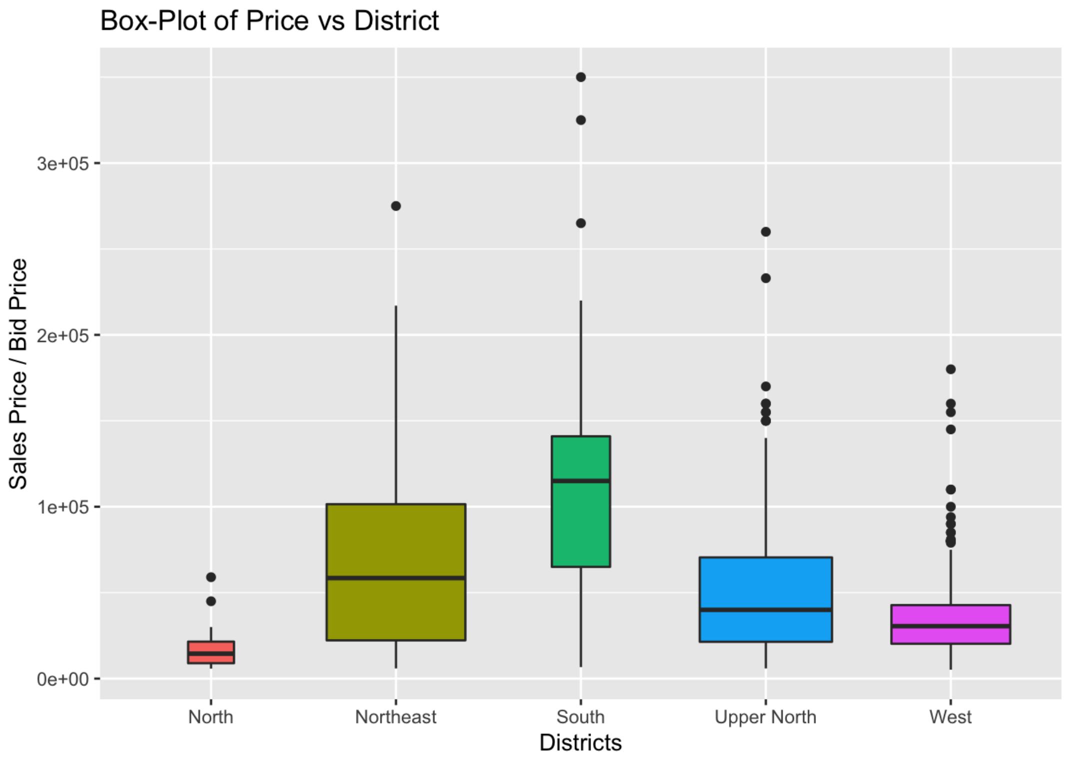


Map based on Longitude (generated) and Latitude (generated). Color shows details about District. The marks are labeled by Average.comps. Details are shown for zip code. The data is filtered on state, which keeps PA.

The map shows the entire Philadelphia County, different color represnets different district, the number gives are the average of final price of different proeprties in differnt postal code.

Now a general view of the data has been provided by the plot shown above. It is obvious that the majority of the plots are located in Northern and Western area of the County. To get a more detailed information regarding each district, please refer to the following box-plot.

```
ggplot(Real_estate, aes(x    = factor(Real_estate$District),
                        y    = Real_estate$Sale.Price.bid.price,
                        fill = factor(Real_estate$District))) +
  geom_boxplot(varwidth = TRUE,
               alpha    = 1.0) +
  theme(legend.position = "none") +
  labs(title = "Box-Plot of Price vs District",
       x     = "Districts",
       y     = "Sales Price / Bid Price")
```

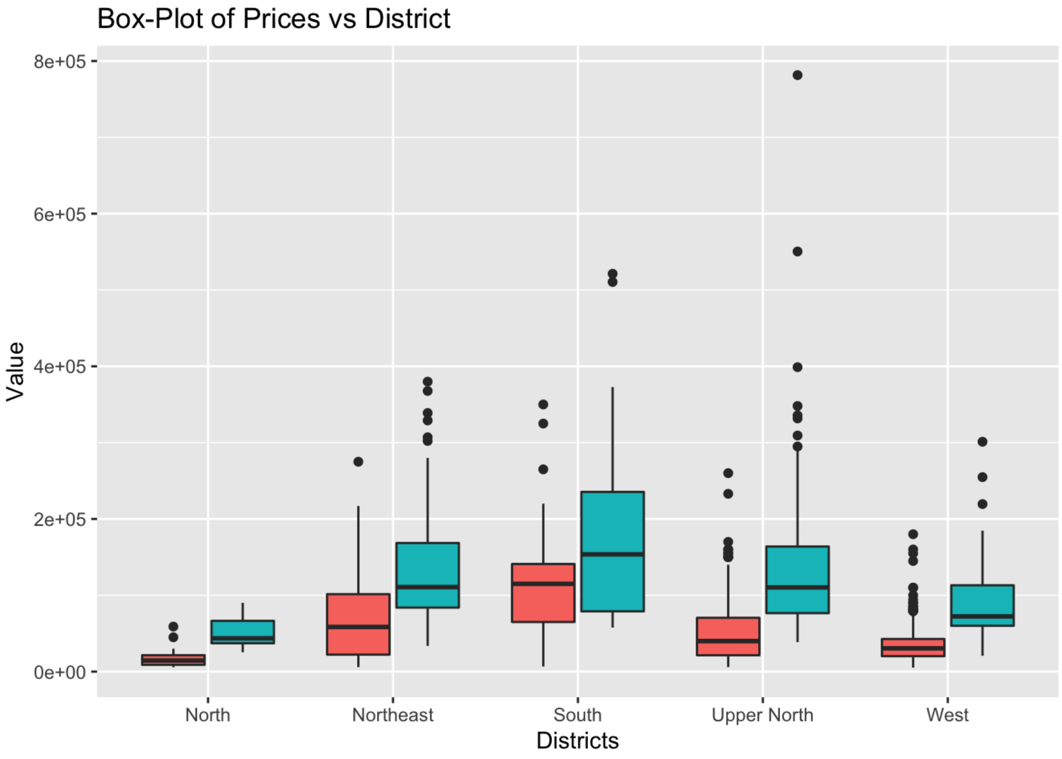


The width of each box refers to the volume of data contained. So this is clear that most of the real estate resources were in auction came from Upper North and Northeast Philly. In fact, these two districts are large in terms of acreage and they are mostly residential area, so they own most of the real estate properties in the Philadelphia County.

North Philly has the lowest average sales price and very few real estate resource. On the other hand, containing a similar volume with North Philly, South Philly has the highest average price and the highest sales price records. This is reasonable because North Philly was one of the oldest area developed in the city so the constructions are old and new constructions tend to favor South Philly better. Thus, these two districts have an opposite sales price trend.

Please note that the sales Price shown above are the closing price from the auction, which is not necessarily representing what the real estate is worth. According to this boxplot, we may conclude that those properties from South Philly are more attractive than others, whereas those from North Philly were the least desired properties. So here comes the question, are the estimated values reflect the same conclusion?

```
compareDataSale  <- data.frame(cbind(Real_estate$Sale.Price.bid.price,
                                     Real_estate$District, "Sale Price"))
compareDataValue <- data.frame(cbind(Real_estate$Zillow.Estimate,
                                     Real_estate$District, "Estimated Value"))
compareData      <- rbind(compareDataSale, compareDataValue)

names(compareData)[1] <- paste("Value")
names(compareData)[2] <- paste("District")
names(compareData)[3] <- paste("Category")

compareData$Value <- as.integer(as.numeric(as.character(sub(",",
                                                            ".",
                                                            compareData$Value))))

ggplot(compareData,
       aes(x     = District,
           y     = Value,
           fill  = Category)) +
  geom_boxplot(alpha    = 1.0) +
  theme(legend.position = "none") +
  labs(title = "Box-Plot of Prices vs District",
       x     = "Districts",
       y     = "Value")
```



In the double boxplot shown above, the green boxes represent estimated value of the properties by Zillow, and the red boxes represent the closing price from auctions. It is clear that the estimated values of the properties are much higher than the sales price in auction. This is normal because this is the reality in the auction industry, which is consumer goods or highly available goods are always sold significantly cheaper than its estimated value, unless it is extremely scarce such as the authentic paintings from Vincent Van Gogh.

According to the comparisons, it is proved that Northeast and South Philly were the most favored by the buyers. The most distinguishable factor is the differences between the auction sales price and the estimated estimated value of the properties in each district. In North, Upper North, and West Philly, 75% of the real estate sold by auction were cheapter than the top 75% of their estimated value in each district. The fact that the sales price from auction being close to the estimated value is a reflection of their popularity, because people were willing to pay for those real estate even though the products were not very much discounted.

Moreover, as for West Philly, though it has the Philadelphia International Airport (PHL) and the 30th Street Station, which are two of the busiest transportation hubs in the area, the strategy of development is highly affected by its special functionality of education. Two of the world’s top universities, University of Pennsylvania and Drexel University, have been neighboring the 30th Street Station for decades. Thus, the demand of real estate in the district would remain relatively stable comparing with other districts, and so is their values.

The analyst decided to split the real estate into categitories that each contains two decades of period, to discover more details about this data.

```r
checkYearBuilt <- function(yearBuilt){
  if(yearBuilt > 2000){
    return("2001-present")
  }
  else if(yearBuilt > 1980){
    return("1981-2000")
  }
  else if(yearBuilt > 1960){
    return("1961-1980")
  }
  else if(yearBuilt > 1940){
    return("1941-1960")
  }
  else if(yearBuilt > 1920){
    return("1921-1940")
  }
  else if(yearBuilt > 1900){
    return("1901-1920")
  }
  else if(yearBuilt > 1880){
    return("1881-1900")
  }
  else if(yearBuilt > 1860){
    return("1861-1880")
  }
  else{
    return(NA)
  }
}

Real_estate$YearBuilt_Dis <- Real_estate$yearBuilt

for(i in 1:(length(Real_estate$YearBuilt))){
  Real_estate$YearBuilt_Dis[i] <- checkYearBuilt(Real_estate$YearBuilt_Dis[i])
}

# excluding the undefined values
Real_estate <- na.omit(Real_estate)

# seting up a table which has specific information of the estimated unit price of the properties
unitPrice <- cbind(data.frame(Real_estate$Zillow.Estimate),
                   Real_estate$finished...SqFt.,
                   Real_estate$PropType,
                   Real_estate$District,
                   Real_estate$YearBuilt_Dis,
                              Real_estate$Zillow.Estimate/Real_estate$finished...SqFt.)

names(unitPrice)[1] <- paste("Zillow.Estimate")
names(unitPrice)[2] <- paste("finished...SqFt.")
names(unitPrice)[3] <- paste("PropType")
names(unitPrice)[4] <- paste("District")
names(unitPrice)[5] <- paste("YearBuilt_Dis")
names(unitPrice)[6] <- paste("Unit.Price.USD/SqFt")

# checking the quality of this table
summary(unitPrice)
```

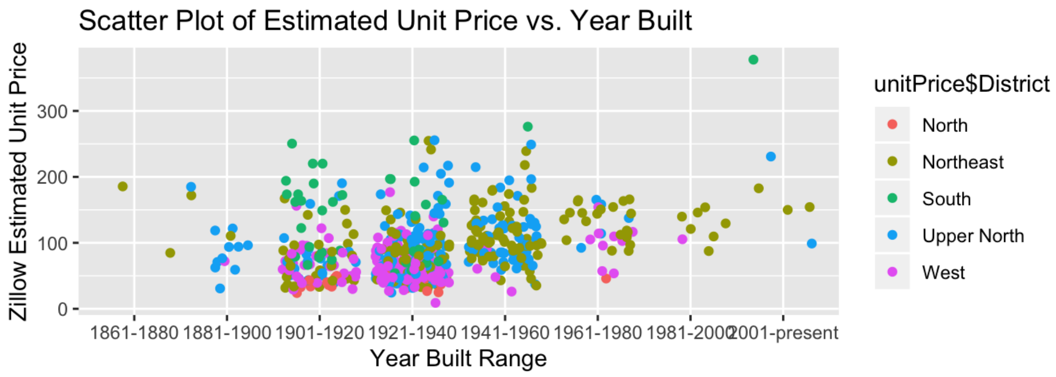
```
##  Zillow.Estimate  finished...SqFt.            PropType          District  
##  Min.   : 20806   Min.   :   0     Condominium     :181   North      : 23  
##  1st Qu.: 66111   1st Qu.:1056     MultiFamily2To4:  5   Northeast  :207  
##  Median : 97724   Median :1200     SingleFamily    :183   South      : 37  
##  Mean   :120186   Mean   :1309     Townhouse       :233   Upper North:187  
##  3rd Qu.:155854   3rd Qu.:1440                            West       :148  
##  Max.   :781509   Max.   :4564                                             
##                                                                            
##    YearBuilt_Dis Unit.Price.USD/SqFt
##  1921-1940:297   Min.   :  8.987    
##  1941-1960:123   1st Qu.: 55.735    
##  1901-1920:117   Median : 82.422    
##  1961-1980: 35   Mean   :    Inf    
##  1881-1900: 14   3rd Qu.:113.162    
##  1981-2000:  8   Max.   :    Inf    
##  (Other)  :  8
```

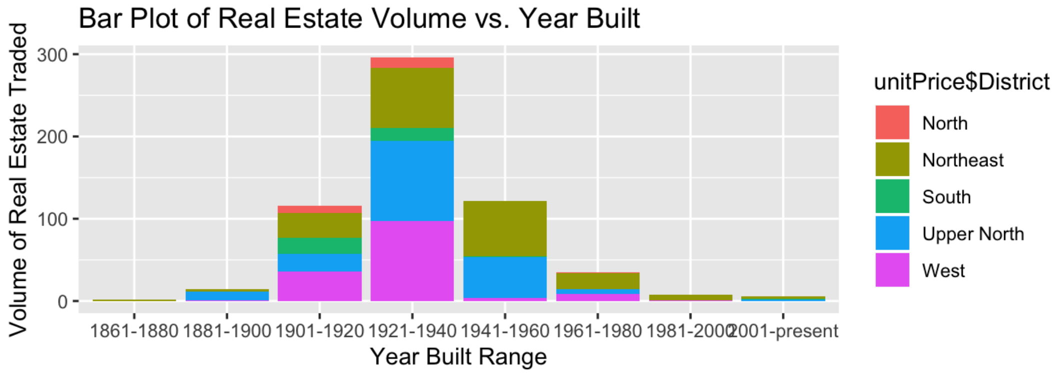
```
# looking for outliers
boxplot.stats(unitPrice$`Unit.Price.USD/SqFt`)
```

```
## $stats
## [1]   8.987473  55.709302  82.422249 113.165116 196.894397
## 
## $n
## [1] 602
## 
## $conf
## [1] 78.72233 86.12217
## 
## $out
##  [1]   254.6359   214.7790        Inf   255.5357   239.1111        Inf
##  [7]   217.0815   214.3210   250.4383 89943.0000   249.0182   204.7280
## [13]   230.8310   220.0551   276.1814   217.9721   255.2560   220.1336
## [19]   377.7319   241.4556
```

So from the data shown above, outliers exist because there were infinity values in the unit price column, which were caused by zero-values in the column of “finished…SqFt.” Even though the calculation decides there are many outliers in the data, the analyst still decide there are only 4 outliers of 3 Inf-values and a 89943.0000 value, because it is possible that the unit price for different property types could be significantly various. To exclude these outliers, the following steps are followed.

```
unitPrice$`Unit.Price.USD/SqFt`[unitPrice$`Unit.Price.USD/SqFt` > 400] <- NA
unitPrice <- na.omit(unitPrice)

library(grid)
library(gridExtra)

graph1 <-
  ggplot(unitPrice,
         aes(x      = unitPrice$YearBuilt_Dis,
             y      = unitPrice$`Unit.Price.USD/SqFt`,
             colour = unitPrice$District)
         ) +
  geom_jitter() +
  labs(title = "Scatter Plot of Estimated Unit Price vs. Year Built",
       x     = "Year Built Range",
       y     = "Zillow Estimated Unit Price")

graph2 <-
  ggplot(unitPrice,
         aes(x    = unitPrice$YearBuilt_Dis,
             fill = unitPrice$District)
       ) +
  geom_bar() +
  labs(title = "Bar Plot of Real Estate Volume vs. Year Built",
       x     = "Year Built Range",
       y     = "Volume of Real Estate Traded")

grid.arrange(graph1, graph2, nrow = 2)
```





The vast majority of these real eastate properties traded in auction were built during the period of 1901-1960, and they are mostly located in Northeast and Upper North Philly. These two graphs, also shows that the Northeast Philly was gradually developing since 1901, and the real estate condition overall is relatively better than those in Upper North Philly, which was highly developed during this period because there is rare during 1881 and 1960. Upper North Philly was pretty much developed as of 1980, because the real estate in the market built after 1980 is rare.

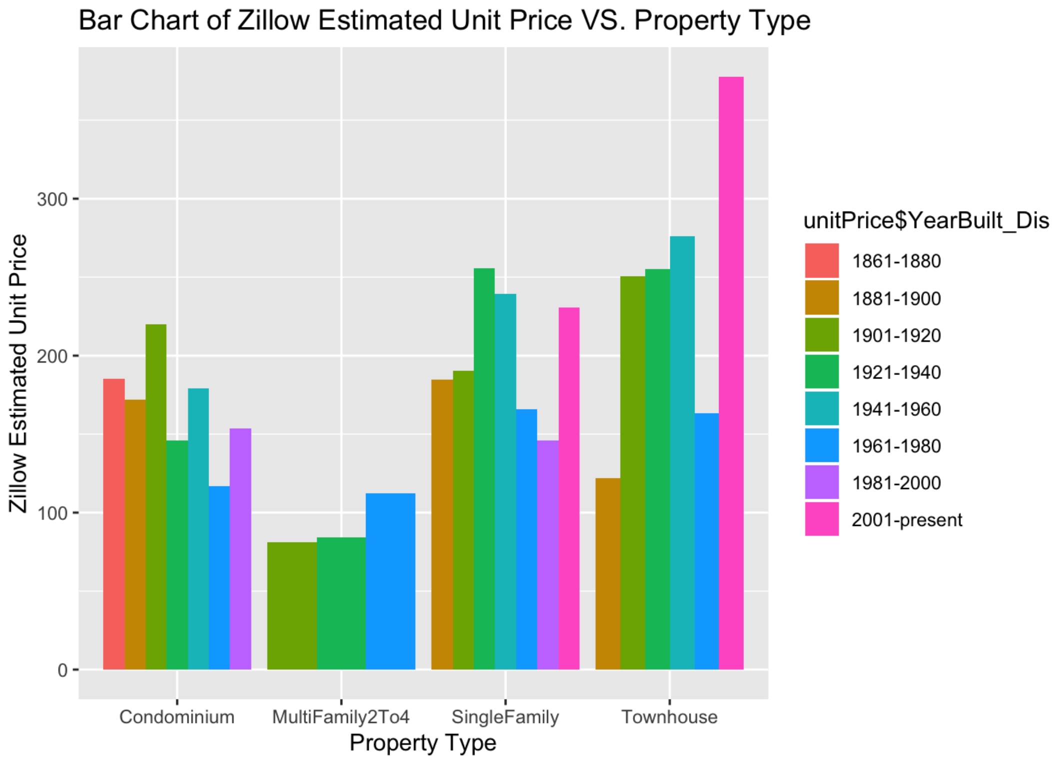
The next questions, who were trading these real estate resources?

```
sellers <- as.character(Real_estate$Seller)

for(i in 1:length(sellers)){
  sellers[i] <- substr(sellers[i], 1, 16)
}

sellers <- data.frame(head(as.table(summary(factor(sellers))),
                           n = 20))
names(sellers)[1] <- paste("Sellers")
names(sellers)[2] <- paste("# Sold")

buyers <- as.character(Real_estate$Buyer)

for(i in 1:length(sellers)){
  buyers[i] <- substr(buyers[i], 1, 16)
}

buyers <- data.frame(head(as.table(summary(factor(buyers))),
                          n = 20))
names(buyers)[1] <- paste("buyers")
names(buyers)[2] <- paste("# Bought")

cbind(sellers, buyers)
```

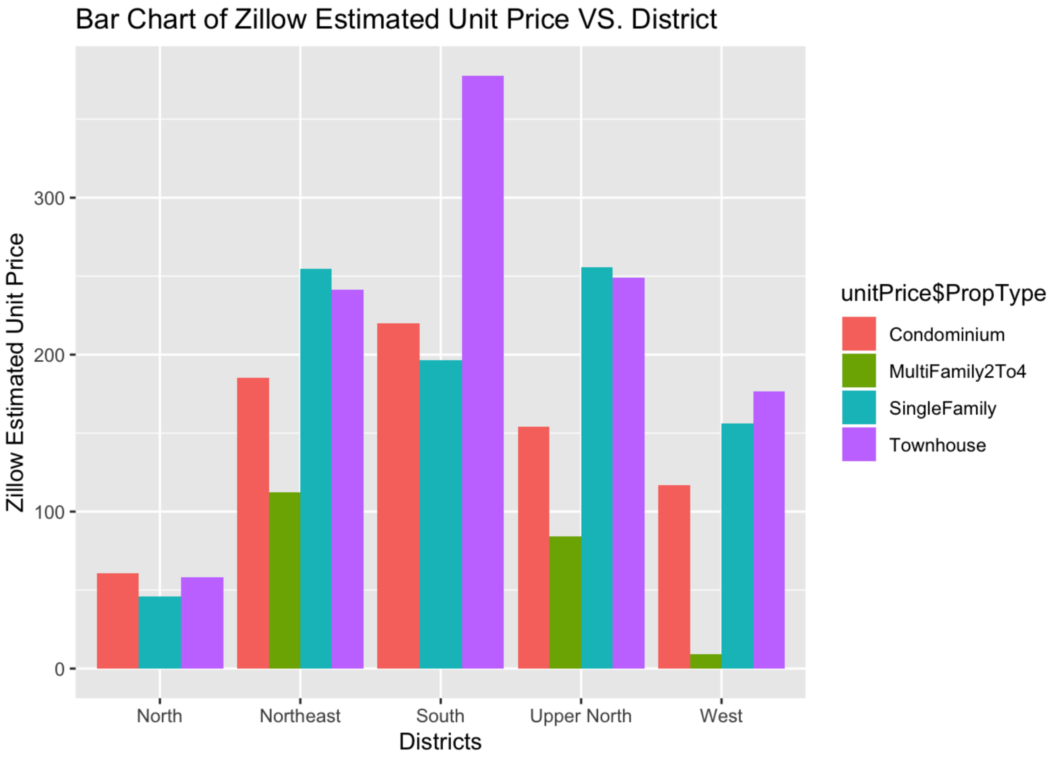
```
##              Sellers # Sold                                 buyers # Bought
## 1  WELLS FARGO BANK     88                   PHELAN HALLINAN  LLP       63
## 2  U.S. BANK NATION     41                          KML LAW GROUP       37
## 3  THE BANK OF NEW      34  FEDERAL NATIONAL MORTGAGE ASSOCIATION       32
## 4  FEDERAL NATIONAL     28                UDREN LAW OFFICES  P.C.       26
## 5  NATIONSTAR MORTG     23              MANLEY DEAS KOCHALSKI LLC       22
## 6  US BANK NATIONAL     22                 SHAPIRO & DENARDO  LLC       18
## 7  BANK OF AMERICA      21        MCCABE  WEISBERG & CONWAY  P.C.       17
## 8  MTGLQ INVESTORS      21              MILSTEAD & ASSOCIATES LLC       16
## 9  WILMINGTON SAVIN     20                STERN & EISENBERG  P.C.       13
## 10 DEUTSCHE BANK NA     19                             COBA  INC.        9
## 11 LSF9 MASTER PART     17                             FANNIE MAE        7
## 12 HSBC BANK USA  N     14                           JOSHUA COHEN        7
## 13 JPMORGAN CHASE B     10            PHILLY 4 SALE BY OWNER  LLC        7
## 14 OCWEN LOAN SERVI     10         POWERS  KIRN & ASSOCIATES  LLC        7
## 15 BAYVIEW LOAN SER      9                     AN PROPERTIES  LLC        6
## 16 NATIONSTAR HECM       9                   CORESTATES GROUP LLC        6
## 17 PNC BANK  NATION      9 FEDERAL HOME LOAN MORTGAGE CORPORATION        6
## 18 REVERSE MORTGAGE      9       HLADIK ONORATO AND FEDERMAN  LLP        5
## 19 SANTANDER BANK        9                LACHALL  COHEN & SAGNOR        5
## 20 CITIMORTGAGE  IN      8                   KSTAR PROPERTIES LLC        4
```

The tables show that the all top trading parties were business. More specifically, the sellers were banks and investment companies; the buyers were mostly law firms and small business. The reason why the banks were massively selling their real estate properties seems interesting to be explored. Since the property types are clear, it is less likely that these sellers intentionally bought the properties in the beginning. The fact that the buyers were small businesses and private persons has proved that these properties were not attractive to the banks or investment companies. Therefore, a reasonable guess could possibly be that these traded properties were mortgages. Due to the incapability of the borrowers paying for their loans, the crediters, which happended to be the banks or investment companies sold these mortgages through auction to minimize loss. Thus, those developing business who were trying to expand their services could find very pleasant deals through auction.

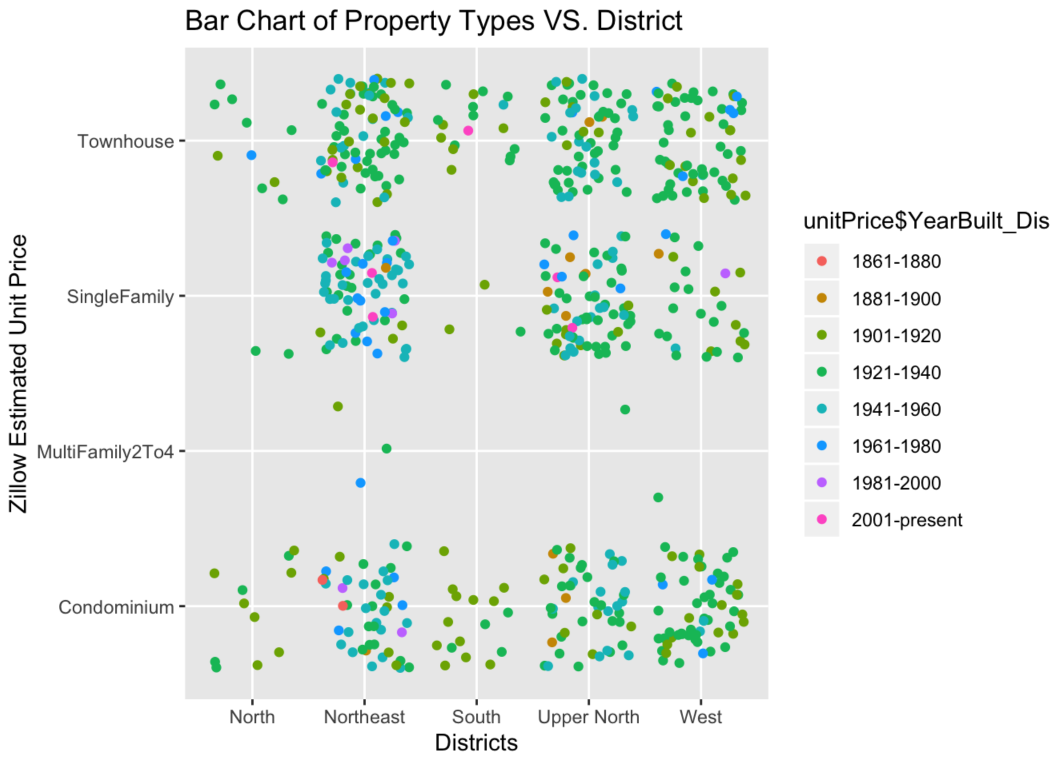
```
ggplot(unitPrice,
       aes(x     = unitPrice$PropType,
           y     = unitPrice$`Unit.Price.USD/SqFt`,
           fill = unitPrice$YearBuilt_Dis))                                   +
  geom_bar(position =  "dodge",
           stat      = "identity")                                            +
  labs(title = "Bar Chart of Zillow Estimated Unit Price VS. Property Type",
       x     = "Property Type",
       y     = "Zillow Estimated Unit Price")
```



```
ggplot(unitPrice,
       aes(x    = unitPrice$District,
           y    = unitPrice$`Unit.Price.USD/SqFt`,
           fill = unitPrice$PropType))                                       +
  geom_bar(position =  "dodge",
           stat     = "identity")                                            +
  labs(title = "Bar Chart of Zillow Estimated Unit Price VS. District",
       x     = "Districts",
       y     = "Zillow Estimated Unit Price")
```



```
library(ggExtra)

ggplot(unitPrice,
       aes(x = unitPrice$District,
           y = unitPrice$PropType,
           color = unitPrice$YearBuilt_Dis))              +
  geom_jitter()                                           +
  labs(title = "Bar Chart of Property Types VS. District",
       x     = "Districts",
       y     = "Zillow Estimated Unit Price")
```



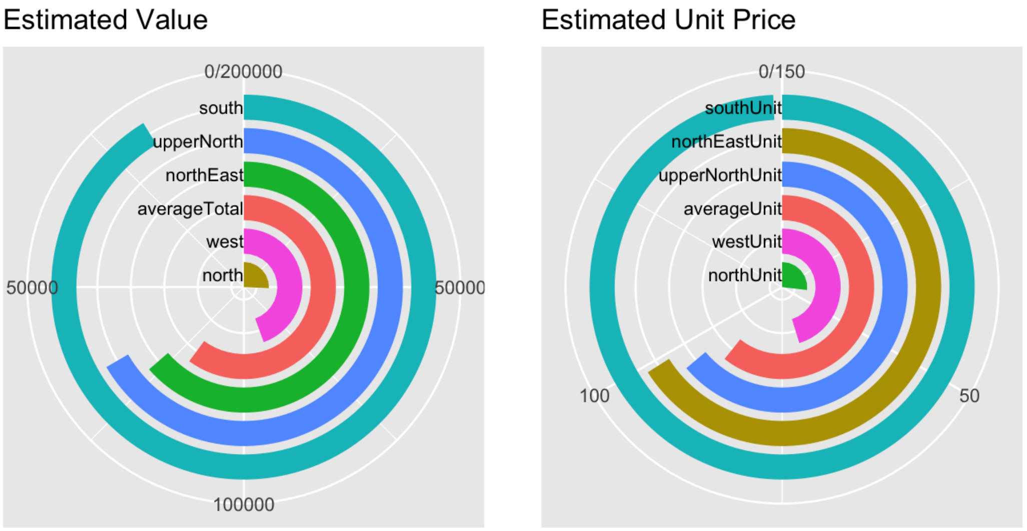
The three graphs above show relationships between the property unit price, the types of these properties, the locations of these properties, and the construction period of these peperties. There are abundant information could be retrieved from these graphs.

Out of Condominium, MultiFamily, SingeFamily, and Townhouse, four types of real estate, Townhouse had the highest average unit price and the most popular throughout most districts, whereas MultiFamily was the opposit against Townhouse. This a reflection of the urban plan of the county. Most of the structures in the area are in fact Townhouse and SingleFamily. So in the scattor plot, the dots are more compact under Townhouse and SingleFamily, except the North Philly, where most of its land are occupied by business and local government. The West Philly tend to have a balanced number of Condominium and Townhouse. This supports the previous statement of its educational function, as it residents are mostly consist of university faculty members and students.

The average unit price in South Philly is much higher than other districts, which also matches with our previous boxplot. MultiFamily, which all were built before 1940, even though it is short on supply, due to its least popularity, it was the least favorable type of property. So, it is significantly cheaper than other types.

Here is a comparison of average sales prices of real estate between districts and the total average price of the County.

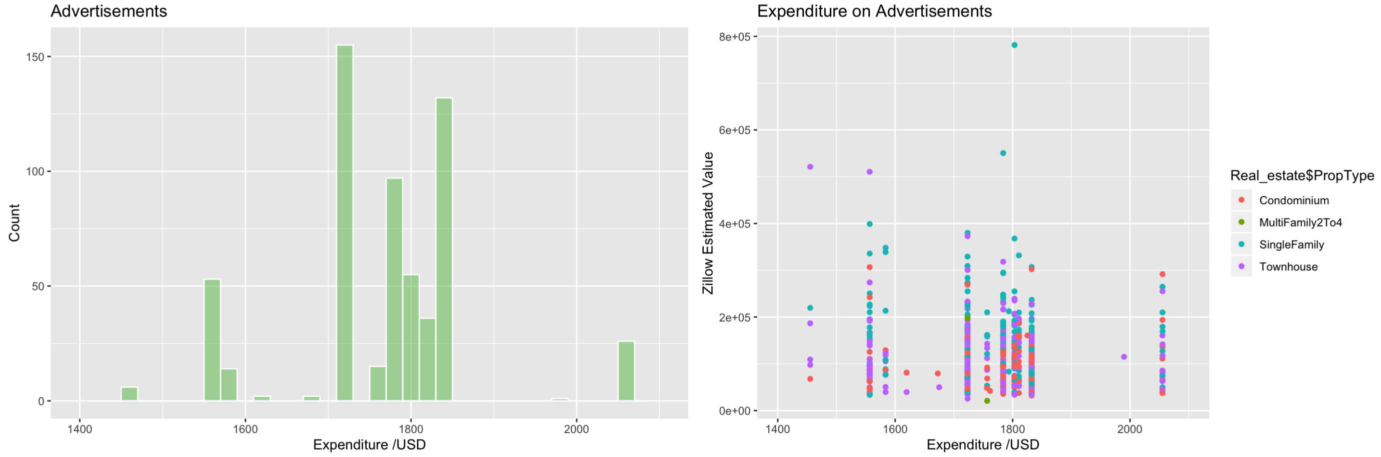
```
averageTotal <- summary(Real_estate$Zillow.Estimate)

north       <- summary(Real_estate$Zillow.Estimate
                      [which(Real_estate$District == "North")])
upperNorth <- summary(Real_estate$Zillow.Estimate
                      [which(Real_estate$District == "Upper North")])
northEast  <- summary(Real_estate$Zillow.Estimate
                      [which(Real_estate$District == "Northeast")])
south       <- summary(Real_estate$Zillow.Estimate
                      [which(Real_estate$District == "South")])
west        <- summary(Real_estate$Zillow.Estimate
                      [which(Real_estate$District == "West")])

tab_total <- cbind(averageTotal, north, upperNorth, northEast, south, west)

# CIRCULAR BARPLOT of Zillow Estimated Price
circular1 <-
  ggplot(data.frame(tab_total),
         aes(x = reorder(colnames(data.frame(tab_total)), tab_total[4,]),
             y = tab_total[4,],
             fill = colnames(data.frame(tab_total)))) +
    geom_bar(width = 0.75, stat="identity") +
    labs(title = "Estimated Value",
         x     = "",
         y     = "") +
    coord_polar(theta = "y") +
    ylim(c(0, 200000)) +
    geom_text(data  = data.frame(tab_total),
              hjust = 1,
              size  = 3,
              aes(x     = reorder(colnames(data.frame(tab_total)), tab_total[4,]),
                  y     = 0,
                  label = reorder(colnames(data.frame(tab_total)), tab_total[4,])
                  )
              ) +
    theme(legend.position = "none" , axis.text.y = element_blank() , axis.ticks = element_blank())



averageUnit <- summary(unitPrice$`Unit.Price.USD/SqFt`)

northUnit      <- summary(unitPrice$`Unit.Price.USD/SqFt`
                          [which(unitPrice$District == "North")])
upperNorthUnit <- summary(unitPrice$`Unit.Price.USD/SqFt`
                          [which(unitPrice$District == "Upper North")])
northEastUnit  <- summary(unitPrice$`Unit.Price.USD/SqFt`
                          [which(unitPrice$District == "Northeast")])
southUnit      <- summary(unitPrice$`Unit.Price.USD/SqFt`
                          [which(unitPrice$District == "South")])
westUnit       <- summary(unitPrice$`Unit.Price.USD/SqFt`
                          [which(unitPrice$District == "West")])

tab_unit <- cbind(averageUnit,
                  northUnit,
                  upperNorthUnit,
                  northEastUnit,
                  southUnit,
                  westUnit)

# CIRCULAR BARPLOT of Estimated Unit Price
circular2 <-
  ggplot(data.frame(tab_unit),
         aes(x = reorder(colnames(data.frame(tab_unit)), tab_unit[4,]),
             y = tab_unit[4,],
             fill = colnames(data.frame(tab_unit)))) +
    geom_bar(width = 0.75, stat="identity") +
    labs(title = "Estimated Unit Price",
         x     = "",
         y     = "") +
    coord_polar(theta = "y") +
    ylim(c(0, 150)) +
    geom_text(data  = data.frame(tab_unit),
              hjust = 1,
              size  = 3,
              aes(x     = reorder(colnames(data.frame(tab_unit)),
                                  tab_unit[4,]),
                  y     = 0,
                  label = reorder(colnames(data.frame(tab_unit)),
                                  tab_unit[4,])
                  )
              ) +
    theme(legend.position = "none" , axis.text.y = element_blank() , axis.ticks = element_blank())

grid.arrange(circular1, circular2, ncol = 2)
```



The average estimated value of real estate in North and West Philly are below the County average.

The next step is to evaluate how much expense on advertising was the most cost-effective for sales by auction?

```
g5 = ggplot(Real_estate,
         aes(x = Real_estate$Advertising)) +
    geom_histogram(binwidth = 20,
                   color = "white",
                   fill = rgb(0.2, 0.7, 0.1, 0.4)) +
    xlim(1400, 2100) +
    labs(title = "Advertisements",
         x = "Expenditure /USD",
         y = "Count")

g6 = ggplot(Real_estate,
       aes(x = Real_estate$Advertising,
           y = Real_estate$Zillow.Estimate,
           colour = Real_estate$PropType)) +
  geom_point() +
  xlim(1400, 2100) +
  labs(title = "Expenditure on Advertisements",
       x     = "Expenditure /USD",
       y     = "Zillow Estimated Value")

grid.arrange(g5, g6, ncol = 2)
```



Most of the sellers spent between 1700 USD and 1800 USD to advertise their properties before auctioning, regardless of the property type and their estimated unit value. Therefore, 1750 USD is detemined to be a directional expenditure suggestion for those who would sell their real estate properties through auction in the future.