Group Project Report

Introduction to Modeling with R

[redacted]

The Situation

For our project we choose to analyze eight years of residential property sales in Johnson County, Iowa from 2007 to 2014. The main question we wanted to answer was “where are the hottest areas of residential property sales in Johnson County, Iowa?” Data from multiple sources was extracted, transformed, and loaded into R to provide a complete data set. Our three main data sources were the Johnson County Assessors web site, Johnson County Assessor’s office, and City of Iowa City Assessor’s office. The Johnson County Assessor’s office gave us the Johnson County’s property attributes for the houses sold from 2007 to 2014. The City of Iowa City Assessor’s office data provided the housing sales and property attributes from Iowa City between the years 2007 to 2014. The data provided contains important attributes such as the Parcel ID, Property Class (single-family, multi-family, agricultural, etc.), Property Address, Sale Date, Sale Amount, Assessed Value (at time of sale), along with other property attributes such as the number of bedrooms, bathrooms, square footage, year built, etc. Using these attributes we merged the different data files together in order to get a complete picture of area residential sales trends.

Answering the Question

Our group cleaned and transformed the data into a clean file in order to analyze the sale trends to answer our main question. In our initial project proposal we hypothesized the areas of North Liberty, Coralville, Solon, and Tiffin will show the highest growth and most sales activity, with rural area University Heights, and most of Iowa City experiencing slower growth. Looking over the data our hypothesis was inaccurate. When we combined Johnson County and Iowa City data, we noted that Iowa City, North Liberty, and Coralville were obvious leaders in the number of residential properties sold (see Figure 1 – 'Number of Residential Properties Sold by City (2007 – 2014)). While the cities of North Liberty and Coralville show a slight increase in growth in the last twenty years, Iowa City leads the area in total sales (see Figure 4 – 'Johnson County Home Sales').

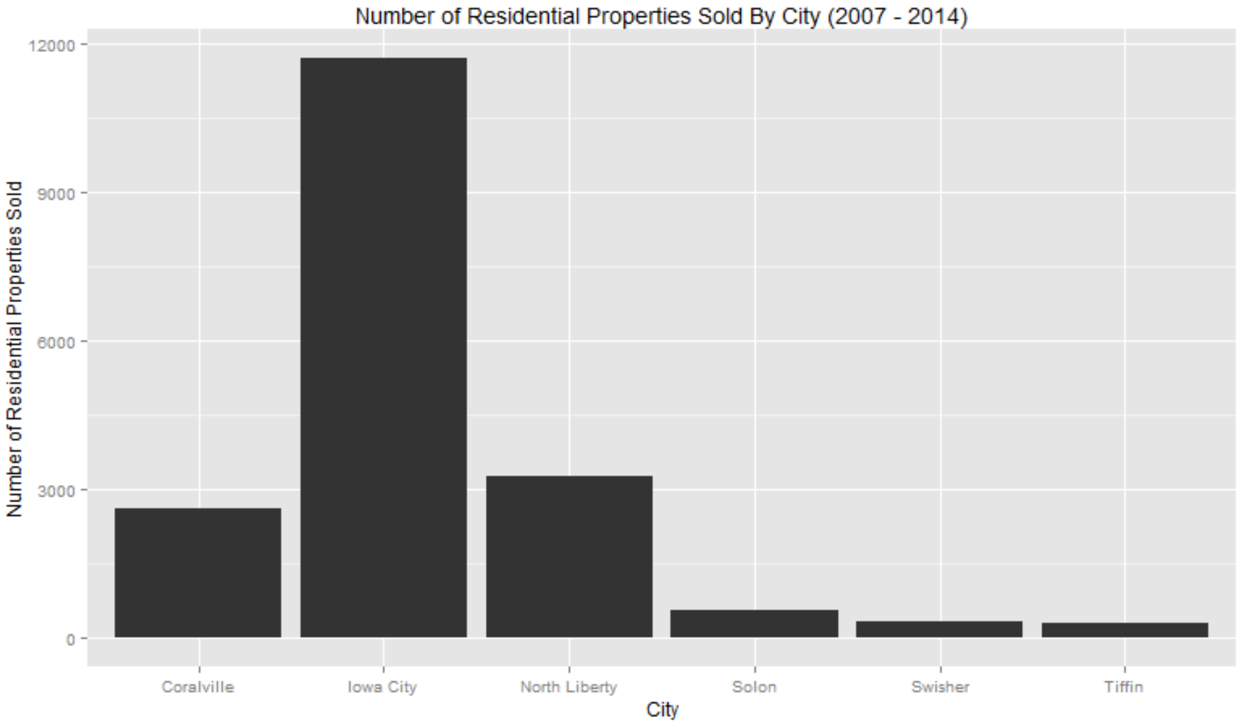


Figure 1

Further Analysis

After we answered the main question our group performed further analysis on the data. We looked at sales according to the following:

- School district
- Age of property
- Occupancy type (renter vs. owner occupied)
- Weekly, monthly, and annual frequency

To perform the school district analysis we used the Johnson County data merged with the district data. The district data contains what school district the houses sold in Johnson County, excluding those sold in Iowa City, belong to. Figure 2, 'Amount of Houses Sold by School District by Year', shows the number of houses sold and the school district they belong to. The school districts Iowa City, Clear Creek, and Solon show the highest number of houses sold in each year.

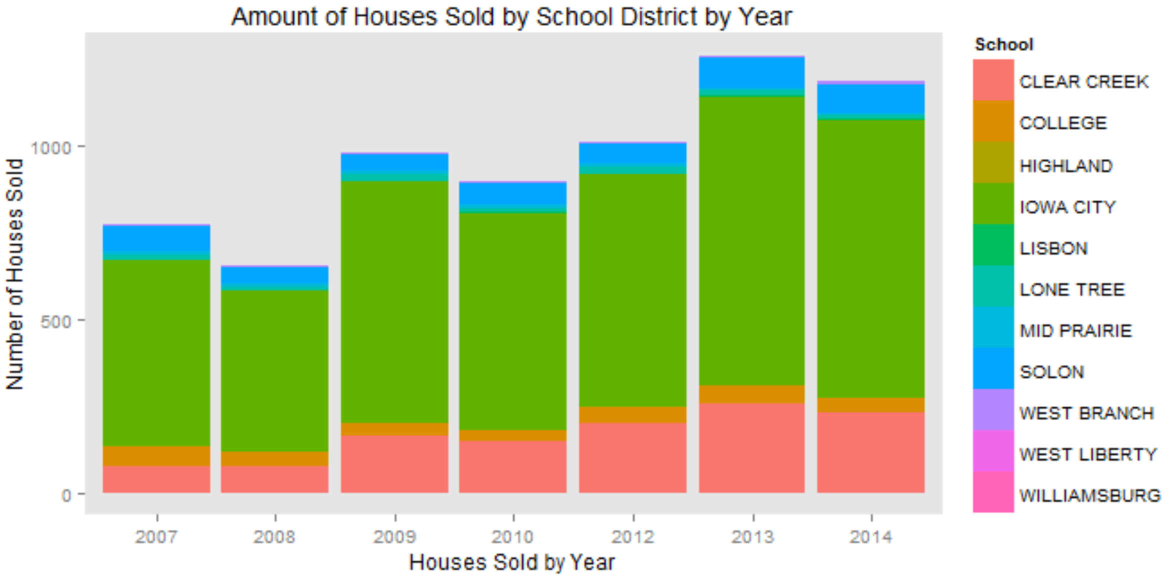


Figure 2

The values of these home sales by school district is demonstrated in Figure 3, ‘School.Sales’. The majority of houses sold for below $250,000 and those which sold for over $250,000 belonged to the Iowa City or Solon school district.

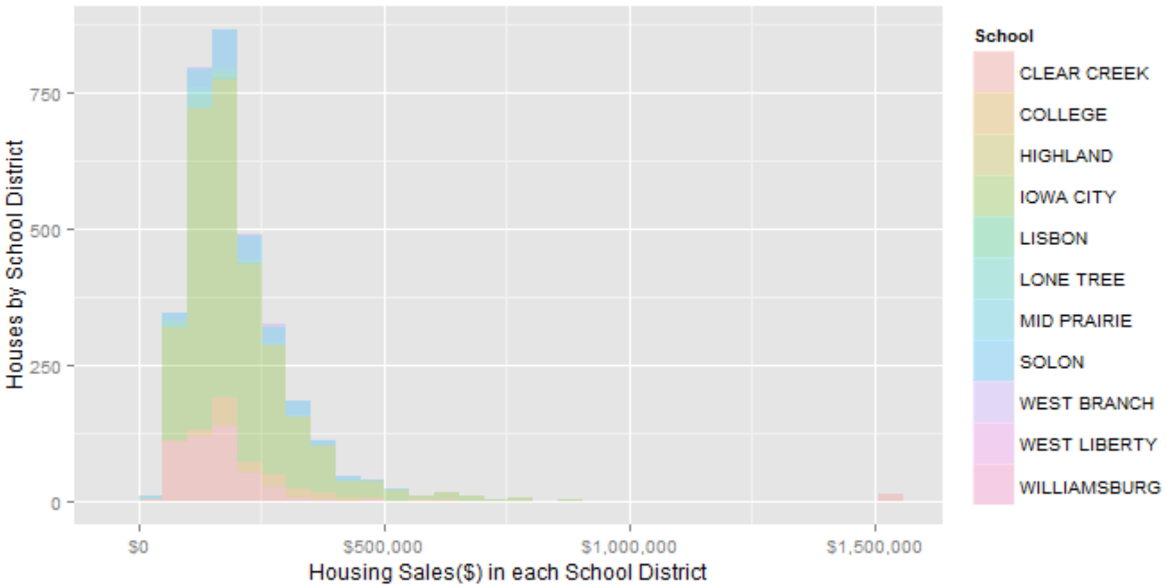


Figure 3

To accomplish a review of data by the age of the home, we created a column, “Age_At_Sale”, using the sales date and built date available in both Johnson County and Iowa City data files. The majority of sales of homes in the top three cities, Iowa City, North Liberty, and Coralville, were homes built in or after the year 2000, revealing these cities as strong areas of new development (Figure 4 – ‘Johnson County Home Sales’).

We also analyzed the occupancy data for the properties sold in Iowa City. We created a column, “Occupancy”, to indicate whether a property is owner occupied or renter occupied by comparing the property address to the mailing address in the file. This information was only available for Iowa City therefore the following analysis does not include the rest of Johnson County. We also focused this analysis on single-family homes and condominiums since these

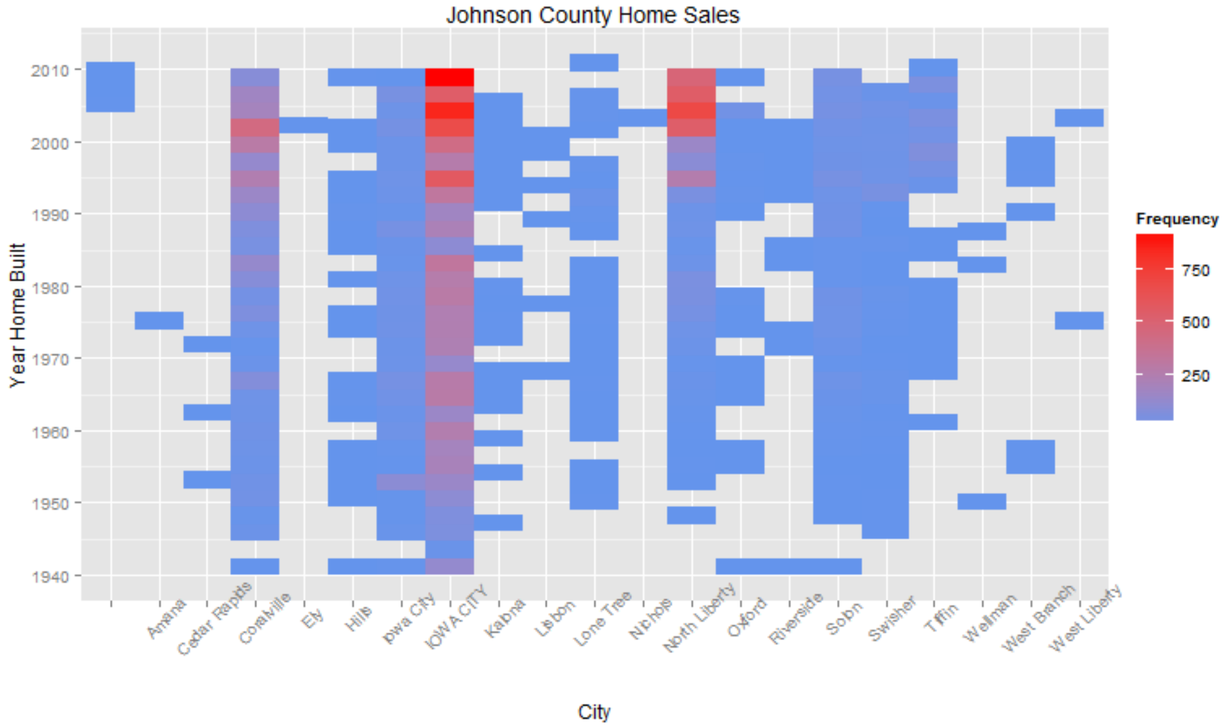


Figure 4

property types make up the majority of Iowa City residential sales. As you can see from Figure 5, 'Sale of Properties by Occupancy Type', renter occupied property sales make up a third of total Iowa City residential sales. From further analysis of Figure 5, we found that single-family homes account for about 80% of the owner occupied properties whereas the majority of renter occupied properties were condominiums. The majority of residential sales for both owner and renter occupied properties occur for properties with values between $125,000 and $150,000; however, there were more sales of condominiums than single family homes at this price point (see 'Condo and Single-Family Home Values by Occupancy Type' in appendix).

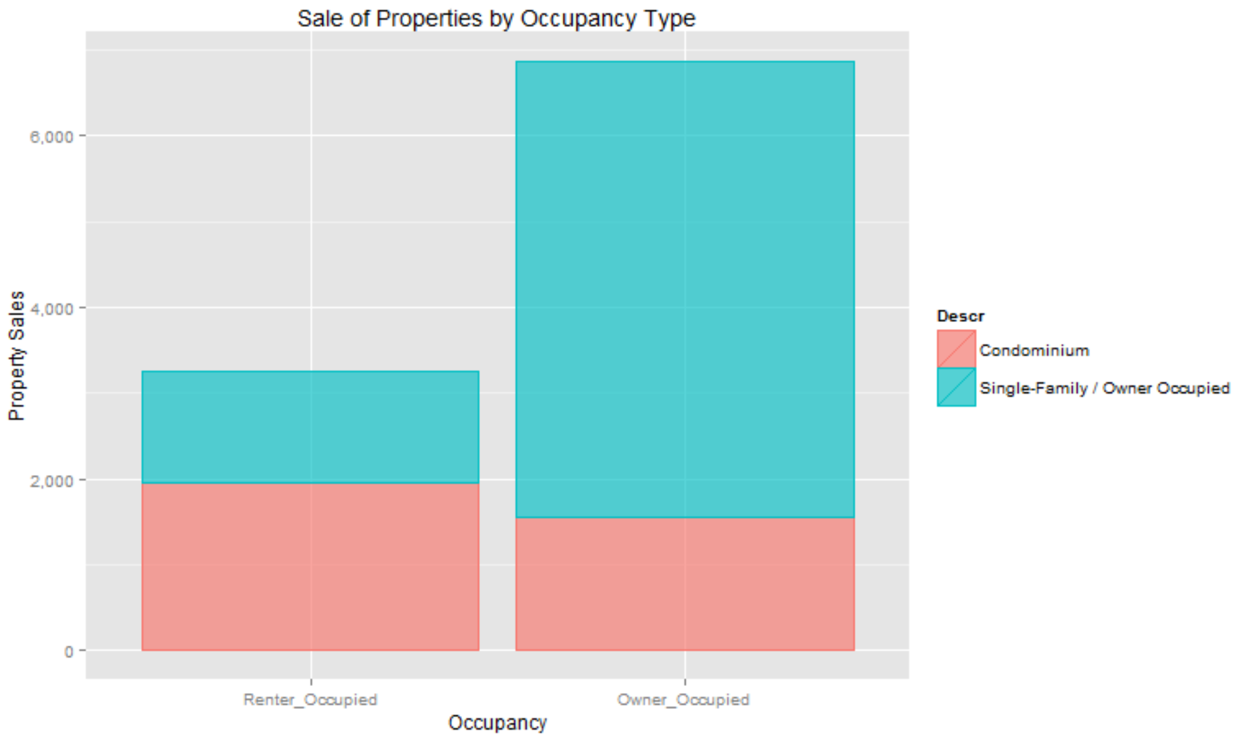


Figure 5

The results of other analysis performed can be viewed in the appendix. A few highlights from this analysis include the following: Our analysis confirmed the theory that summer is the peak season for home sales; Johnson County saw an increase in property sales in 2010 compared to surrounding years; With the exception of a few outliers, sales amounts are aligned with assessed value; Owner occupied properties generally have a higher value than renter occupied properties; and Homes built in 2004 continued to be a major source of sales between 2007 and 2014.

Getting the Data

Residential property sales by year data are publically available from the Johnson County Assessor's web site (http://www.johnson-county.com/dept_jc_assessor.aspx?id=7337), in the form of one Excel spreadsheet per year for a total of eight files. The Johnson County district information (i.e., district codes with associated full name and school district) was originally an HTML table that was copied from the Johnson County Assessor's website and saved as a .csv file. These spreadsheets include basic variables such as Parcel ID, Address, City, Sale Amount, and Assessed Value at the time of the sale. The Johnson County residential property attributes were obtained by visiting the Johnson County Assessor's Office in Iowa City and requesting a report. The City of Iowa City sales and residential property attributes were obtained by visiting the City of Iowa City Assessor's Office (located in the same building as the Johnson County Assessor's Office – in the suite next door). Once we had the data files the files were imported into R as separate data frames, columns were cleaned and aligned, and then all data frames were combined into a single data frame for analysis.

Obstacles Faced

Our group faced several large obstacles in completing this project. Once the project was underway we realized the data we received required a lot of cleaning up and formatting. The majority of our focus for this project was cleaning up the data and extracting the necessary information. The Johnson County and City of Iowa City use different systems to store data, which came in different formats, including .xls, .xlsx, .dbf, and .csv. We wrote code using the 'xlsx' package to read the Excel-formatted files directly, and the read.dbf function from the 'foreign' package to read the .dbf file, but these functions were much slower than importing.csv files using the read.csv function. It took several minutes to import the data using the external packages versus a couple seconds to import data using the read.csv function. We converted all of the original files into .csv files for convenience and performance.

Since the data was comprised of three different data sources and each data source was broken out in years the variables in each file differed from year to year and from source to source. In the eight years of sales data downloaded from the Johnson County Assessor's website there are three different sets of columns available. In other words, the available variables for 2007-2010 are different (in terms of column name, class when imported, and ordering) than what's available for 2011, and 2012-2014 are also different from previous years. The first half of the data set (2007-2010) contains 19 variables, and the 2011-2014 data sets contain 12 variables. Once we dropped the columns that we didn't care about or weren't available across all years we ended up with ten variables for each year before merging with the residential property attribute data. Data from the City of Iowa City was cleaner in this regard, as it came in two spreadsheets (one for sales, and another for residential property attributes). This meant there was less clean up to do on the columns (still need to rename, reclassify, and reorder), but we only had to do it once per column and not once for every column times eight data frames.

One noted obstacle was when we initially got banned from the Beacon Schneidercorp website for scraping data or additional property attributes. When our group was just starting the group project, we wrote a 'for' loop to scrape the Beacon Schneidercorp website for detailed property assessment information to get additional property data. The loop constructed the appropriate property URL by appending the parcel ID to a base URL (i.e., http://beacon.schneidercorp.com/.../[Parcel.ID].html), then it download and parse the elements on the page so we could get things like lot size, tax information, etc. This worked great until running the script all the way through triggered an anti-scraping mechanism that blocked our IP addresses from using the site. We wrote an apology email explaining that we read the site's Terms and Conditions before running the script and didn't see anything about not scraping their site. They removed the IP block and told us not to run scraping scripts, which we agreed to.

To plot property data on a map, which is something we wanted to do to visualize differences in sales activity across different areas of the county, we first needed to get latitude and longitude values for each property address. To do this, we attempted to utilize the RDSTK (R Data Science Toolkit), which is a package with several interesting functions, including street2coordinates which takes a street address and returns latitude and longitude information. Not wanting to get banned from another website, we downloaded the entire Data Science Toolkit website, which is an available option, and ran it as a virtual machine on our computer. This was much better in terms of performance, as the API being accessed is running on our local network and not across the internet. Another option we explored was to use the Google Maps API. This would have worked for us, but Google imposes a rate limit of 1,000 API calls per day, and we had somewhere around 20,000 records to update. In the end, were not successful in importing the longitude and latitude data using these options.

Function Documentation

The functions we wrote for this project are designed to help visually answer the question of hottest areas in terms of residential home sales in Johnson County. One function ("pfunc.Sales.Per.Year.By.City") takes the city name (i.e., "Iowa City", "Coralville", "North Liberty", etc.) as input and returns a histogram of sales in that city by year, figure five. We also dynamically set the main label to include the parameter, add an x-axis label, y-axis label, and appropriately format the axis.

```
# Sales per Year by City
pfunc.Sales.Per.Year.By.City <- function(city = "Tiffin"){
  tempdf <- subset(dfAll, City == city)
  tempp <- qplot(Sale.Year, data = tempdf, geom = "histogram",
                 main = paste("Number of Residential Properties Sold in ", city, sep = ""),
                 xlim = c("2007", "2008", "2009", "2010", "2011", "2012", "2013", "2014"),
                 xlab = "Sale Year",
                 ylab = "Number of Sales")
  tempp <- tempp + scale_y_continuous(labels = comma)
  tempp
}
```

An example of the output of this function using North Liberty as the city input is displayed in Figure 6 – ‘Number of Residential Properties Sold in North Liberty’.

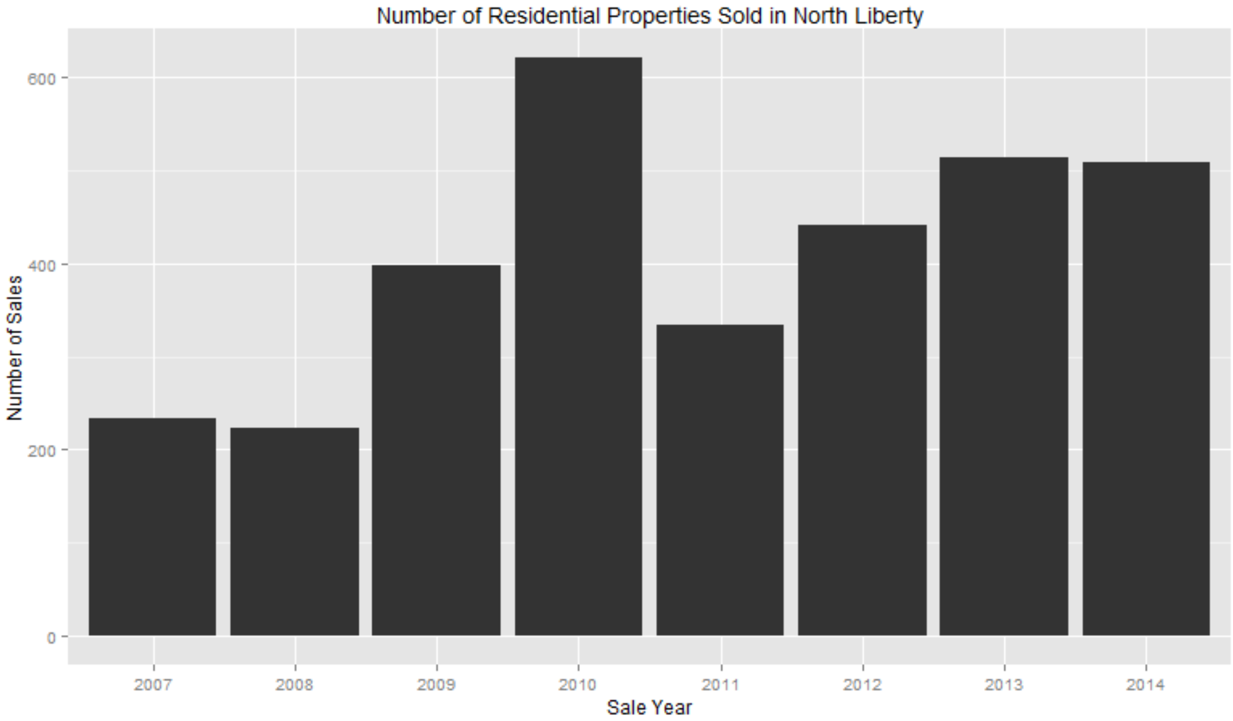


Figure 6

The other function (“pfunc.Sales.Per.Year.By.School”) takes a school district name (i.e., “Iowa City”, “Clear Creek-Amana”, “Solon”, etc.) as input and returns a histogram of sales in that school district by year, figure six. We again dynamically set the main label to include the input parameter, and add formatting to the x- and y-axis.

```
# Sales Per Year By School
pfunc.Sales.Per.Year.By.School <- function(school = "Iowa City"){
  tempdf <- subset(dfAll, School == school)
  tempp <- qplot(Sale.Year, data = tempdf, geom = "histogram",
                 main = paste("Number of Sales in ", school, " School District", sep = ""),
                 xlab = "Sale Year",
                 ylab = "Number of Sales")
  tempp <- tempp + scale_y_continuous(labels = comma)
  tempp}
```

An example of the output of this function is displayed in Figure 7 – ‘Number of Residential Properties Sold in the Clear Creek-Amana School District (2007 – 2014).

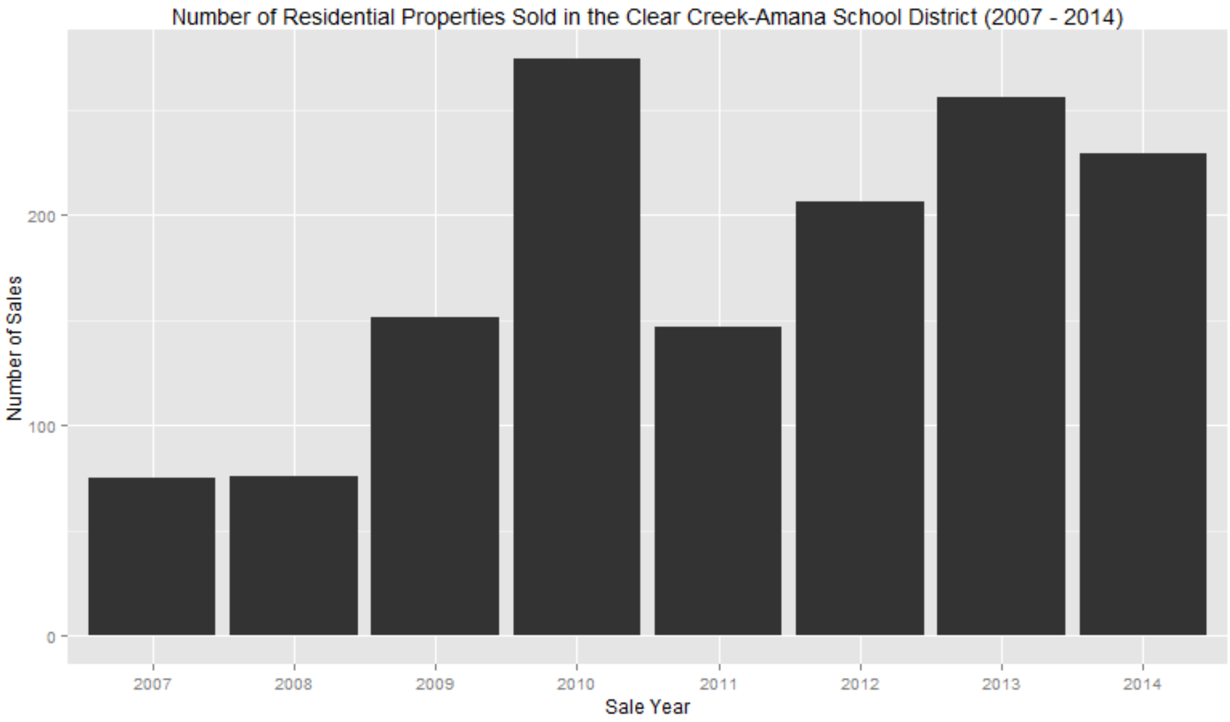
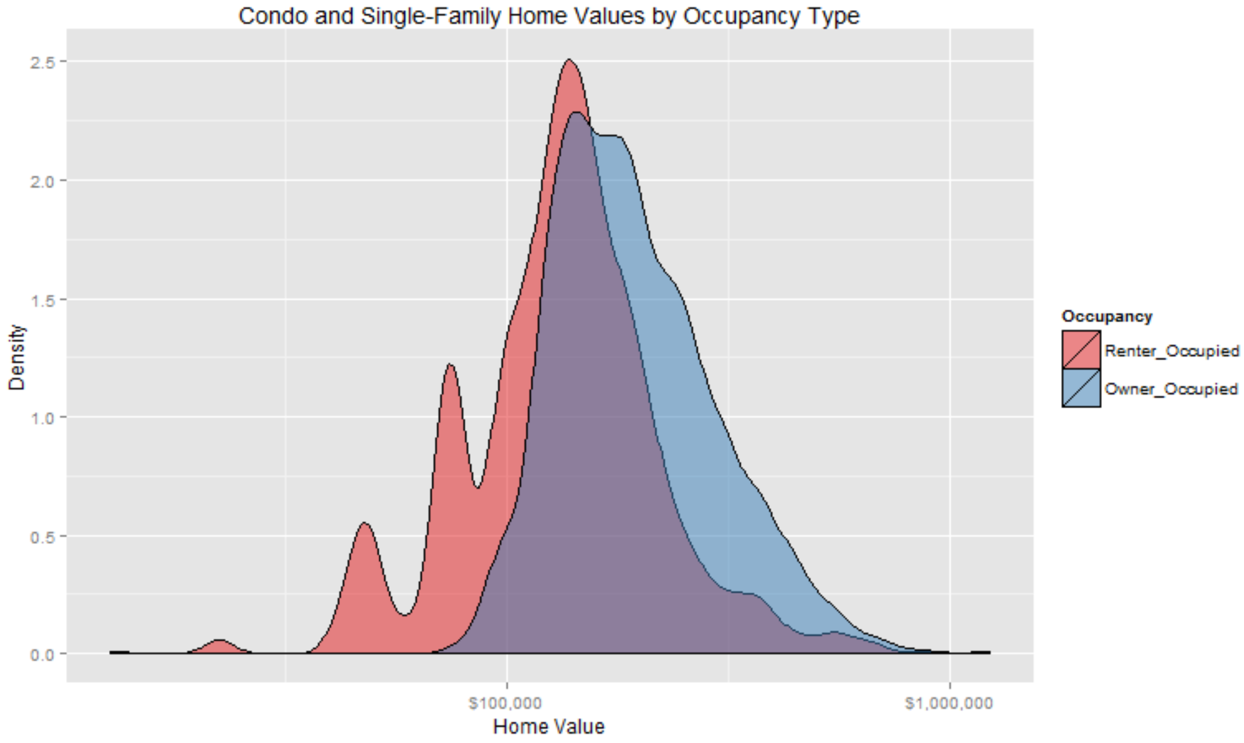
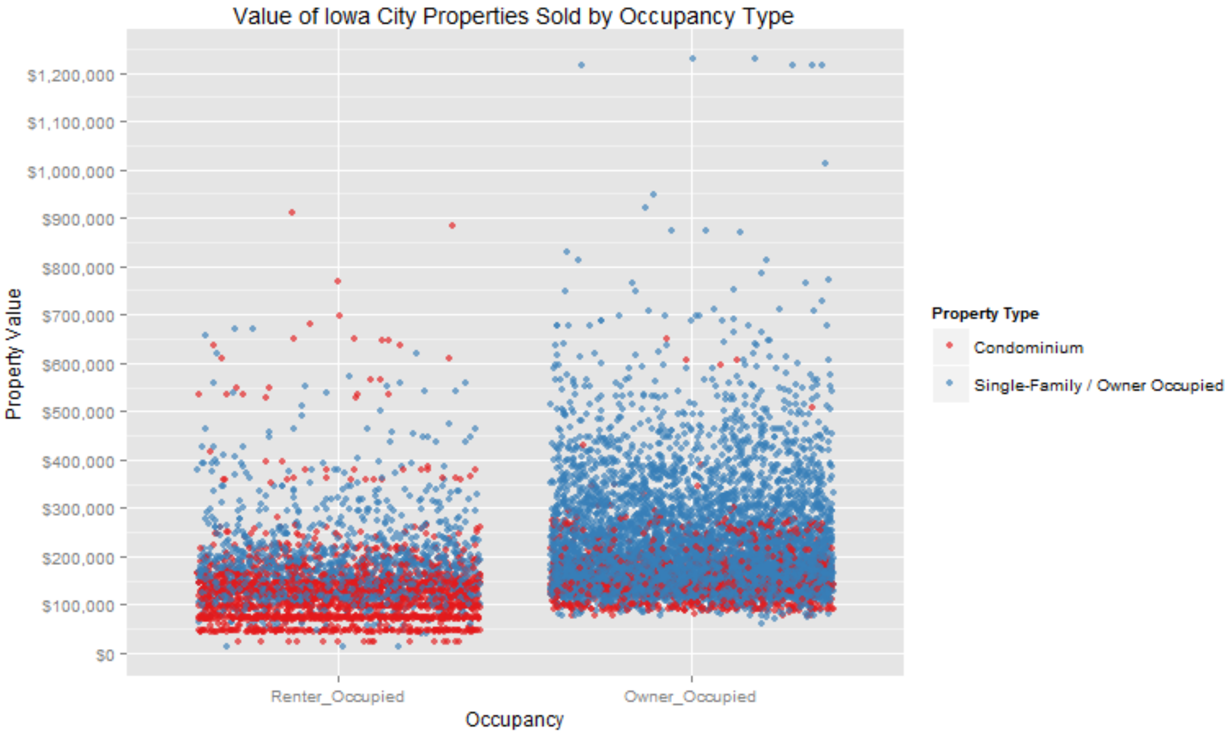
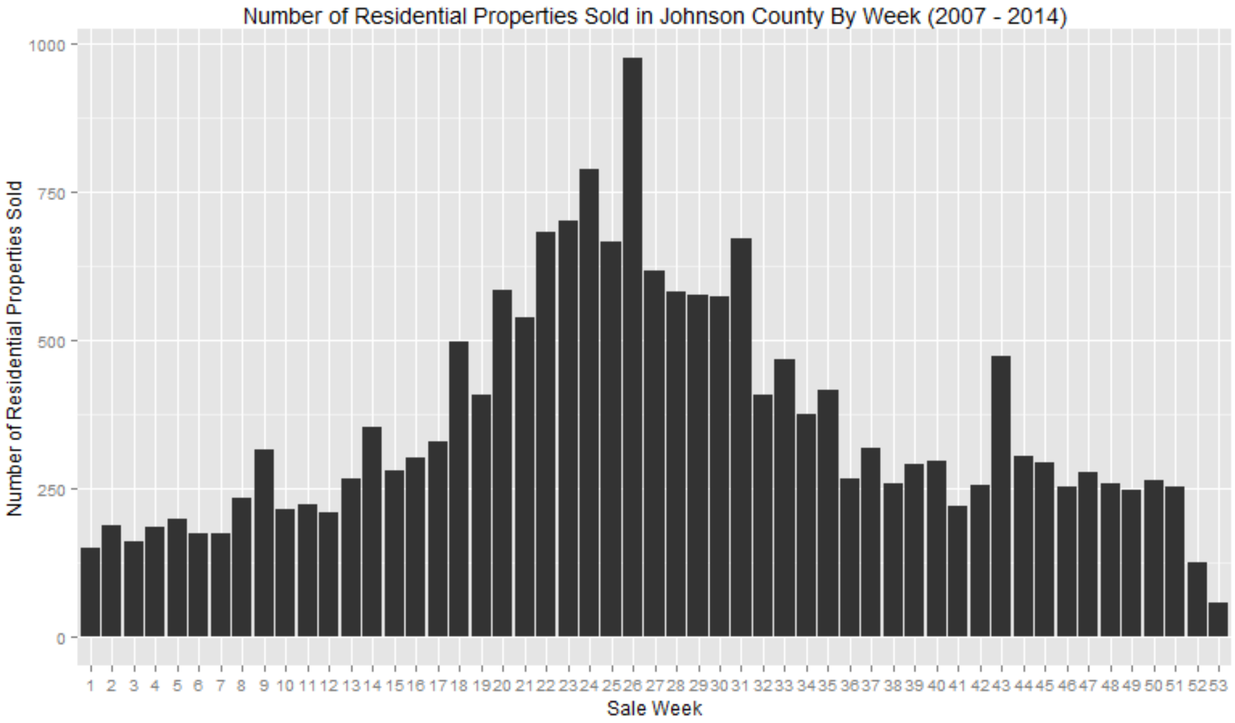


Figure 7

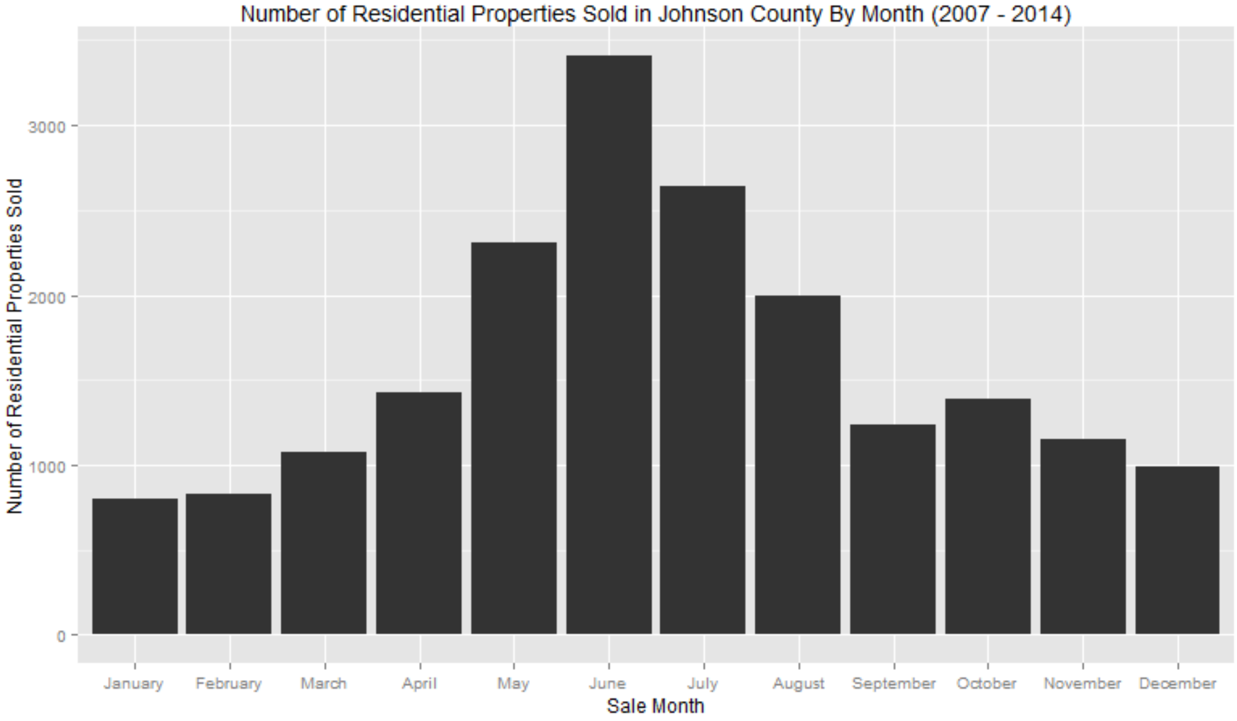
Appendix


Condo and Single-Family Home Values by Occupancy Type
Density
2.5
2.0
1.5
1.0
0.5
0.0
$100,000
$1,000,000
Home Value
Occupancy
Renter_Occupied
Owner_Occupied

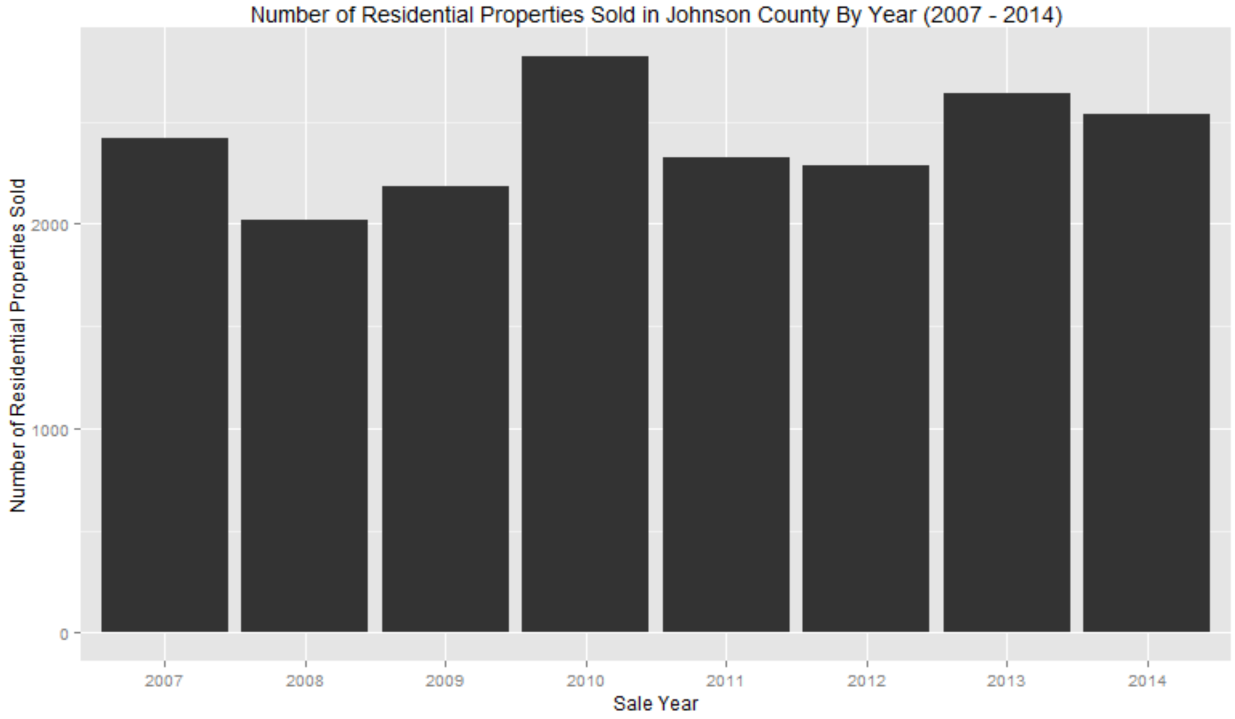

Value of Iowa City Properties Sold by Occupancy Type
Property Value
$1,200,000
$1,100,000
$1,000,000
$900,000
$800,000
$700,000
$600,000
$500,000
$400,000
$300,000
$200,000
$100,000
$0
Renter_Occupied
Owner_Occupied
Occupancy
Property Type
Condominium
Single-Family / Owner Occupied

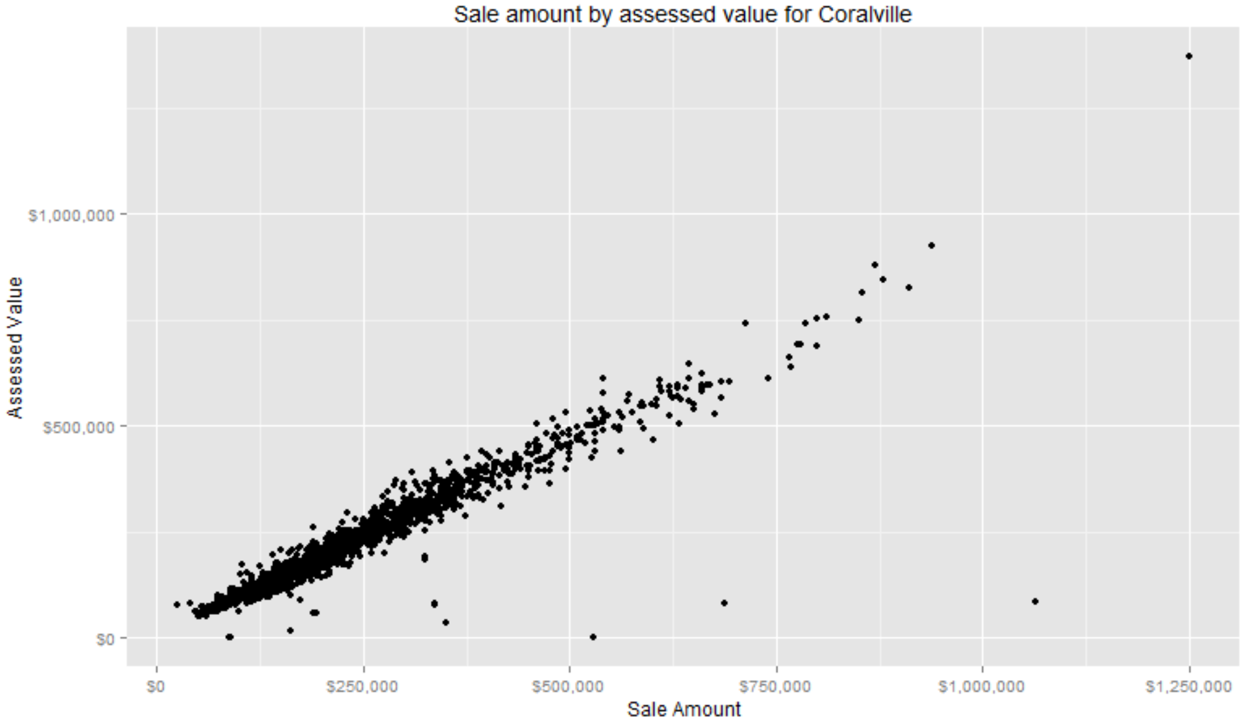


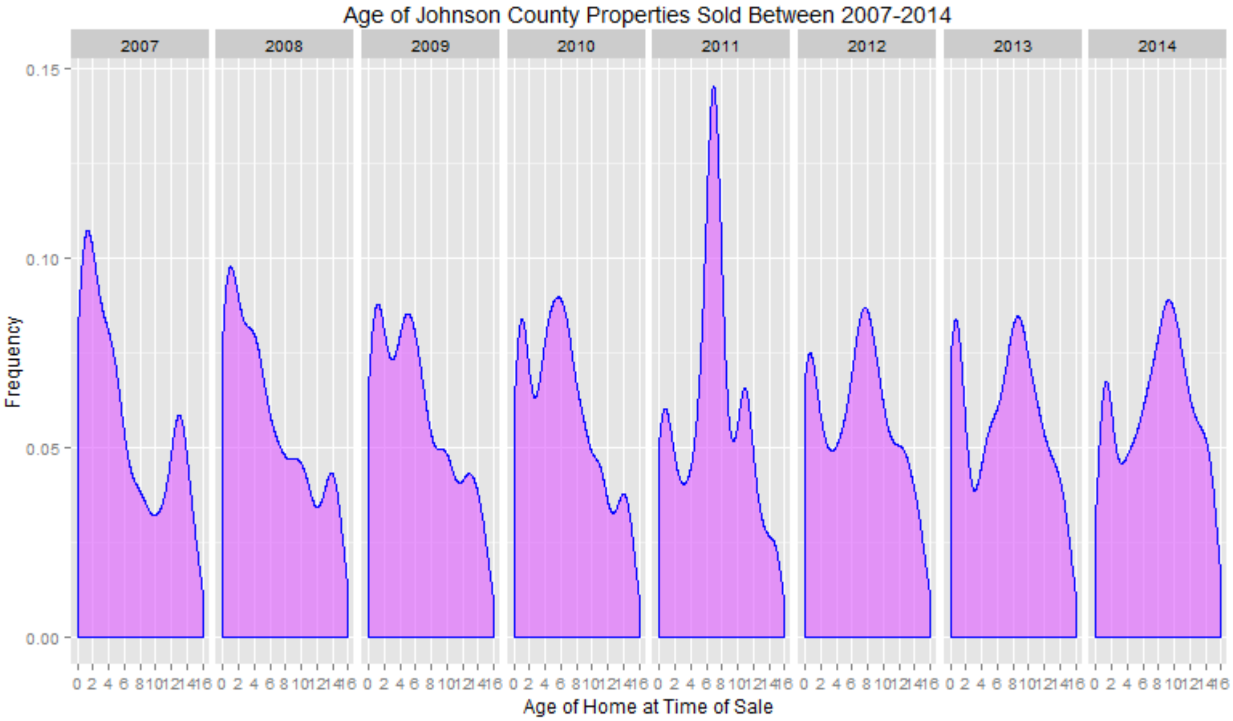
Number of Residential Properties Sold in Johnson County By Week (2007 - 2014)



Number of Residential Properties Sold in Johnson County By Month (2007 - 2014)


Number of Residential Properties Sold in Johnson County By Year (2007 - 2014)
Number of Residential Properties Sold
0
1000
2000
2007
2008
2009
2010
2011
2012
2013
2014
Sale Year


Sale amount by assessed value for Coralville
Assessed Value
$0
$500,000
$1,000,000
$0
$250,000
$500,000
$750,000
$1,000,000
$1,250,000
Sale Amount



Age of Johnson County Properties Sold Between 2007-2014