Module09

Violet Lingenfelter March 22, 2019

Correlation

For looking at correlations, I decided calculate a matrix for all variables I might find interesting in a sort of exploratory approach. I included the following variables in my matrix:

- building, land, and total values
- list price
- number of rooms
- residential area
- building area

This matrix can be seen below.

```
# import package that we need
require(Hmisc)

# look at relationships between a TON of columns
# columns looked at: building, land, other, and total value, lot size, list price, year built, numbe
correlations<-rcorr(as.matrix(assessor[c(4,5,7,11,31,32,35)]))
# class(correlations)


require(formattable)
r <- do.call(rbind.data.frame, correlations[1])

r <- round(r, digits = 3)

# look at r
r <- formattable(r, list(BLDG_VAL = color_tile("white", "#a3dec9"),
                    LAND_VAL = color_tile("white", "#a3dec9"),
                    TOTAL_VAL = color_tile("white", "#a3dec9"),
                    LS_PRICE = color_tile("white", "#a3dec9"),
                    UNITS = color_tile("white", "#a3dec9"),
                    RES_AREA = color_tile("white", "#a3dec9"),
                    NUM_ROOMS = color_tile("white", "#a3dec9")))

r
```

	BLDG_VAL	LAND_VAL	TOTAL_VAL	LS_PRICE	UNITS	RES_AREA	NUM_ROOMS
r.BLDG_VAL	1.000	0.642	0.944	0.497	0.658	0.813	0.788
r.LAND_VAL	0.642	1.000	0.846	0.524	0.685	0.618	0.862
r.TOTAL_VAL	0.944	0.846	1.000	0.551	0.720	0.805	0.827
r.LS_PRICE	0.497	0.524	0.551	1.000	0.529	0.355	0.536
r.UNITS	0.658	0.685	0.720	0.529	1.000	0.513	0.905
r.RES_AREA	0.813	0.618	0.805	0.355	0.513	1.000	0.924
r.NUM_ROOMS	0.788	0.862	0.827	0.536	0.905	0.924	1.000

From looking at this matrix, we can see how correlated some of our variables are with each other. To make this more readily apparent, I chose to highlight each piece of the table based on the value of r. We can see that the least correlated variable to any other variable is LS_PRICE. This means that list prices have little correlation with any of the other variables I chose to include in the matrix.

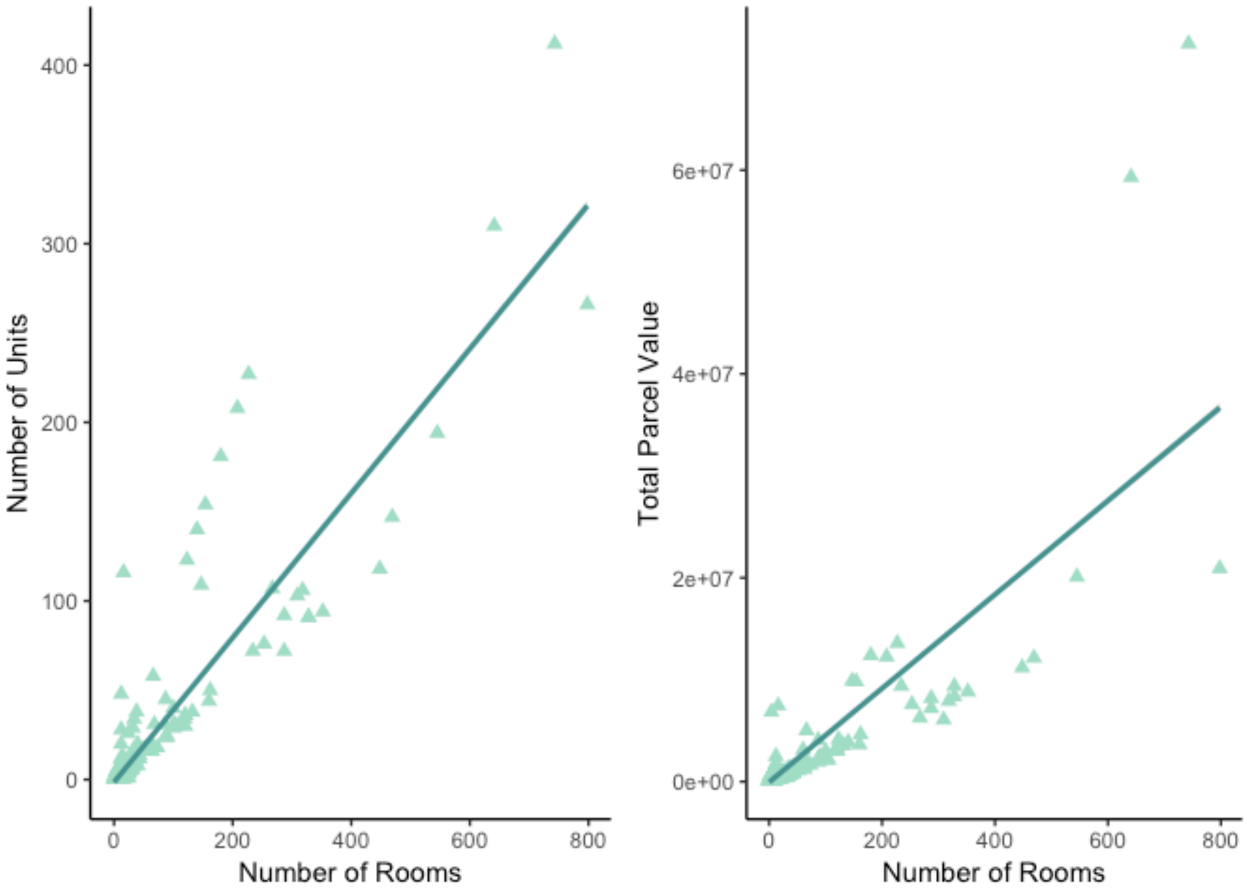
An interesting column to look at is the NUM_ROOMS column. We can see that the number of rooms a building on a parcel had has a strong positive correlation with every variable besides LS PRICE. This is interesting because it is not something I expected.

```
require(ggplot2)
require(cowplot)

# make scatter plot with regression
plot1 = ggplot(assessor, aes(x=NUM_ROOMS, y=UNITS)) +
  theme_classic() +
  geom_point(size=2, color="#a3dec9", shape=17) +
  geom_smooth(method=lm, color="#499491") +
  xlab("Number of Rooms") +
  ylab("Number of Units")

plot2 = ggplot(assessor, aes(x=NUM_ROOMS, y=TOTAL_VAL)) +
  theme_classic() +
  geom_point(size=2, color="#a3dec9", shape=17) +
  geom_smooth(method=lm, color="#499491") +
  xlab("Number of Rooms") +
  ylab("Total Parcel Value")

plot_grid(plot1, plot2)
```



```
regression<-lm(NUM_ROOMS~TOTAL_VAL+UNITS+RES_AREA, data=assessor)

summary(regression)
```

```
## 
## Call:
## lm(formula = NUM_ROOMS ~ TOTAL_VAL + UNITS + RES_AREA, data = assessor)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -235.577   -1.249   -0.074    1.087  250.208 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3.852e+00  5.546e-02   69.46   <2e-16 ***
## TOTAL_VAL   -9.662e-06  1.740e-07  -55.52   <2e-16 ***
## UNITS        6.738e-01  1.928e-02   34.95   <2e-16 ***
## RES_AREA     2.719e-03  3.016e-05   90.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.713 on 12105 degrees of freedom
##   (3000 observations deleted due to missingness)
## Multiple R-squared:  0.8928, Adjusted R-squared:  0.8928 
## F-statistic: 3.36e+04 on 3 and 12105 DF,  p-value: < 2.2e-16
```

```
assessor<-merge(assessor,data.frame(regression$residuals),by='row.names',all.x=TRUE)
```

Regression

We can see from this regression, where we looked at total value, number of units, and residential area as explanatory variables for number of rooms, that our model is okay but missing some variables. We can see that 89% of the variation in number of rooms can be explained by total value, number of units, and residential area. We can map these residuals so as to see if any geographic areas are not explained by this non geographically weighted regression.

```
require(rgdal)
require(sp)

gdb <- "../M248_parcels_gdb/M248_parcels_sde.gdb"

assessor_geo<-readOGR(dsn=gdb, layer="M248TaxPar")
```

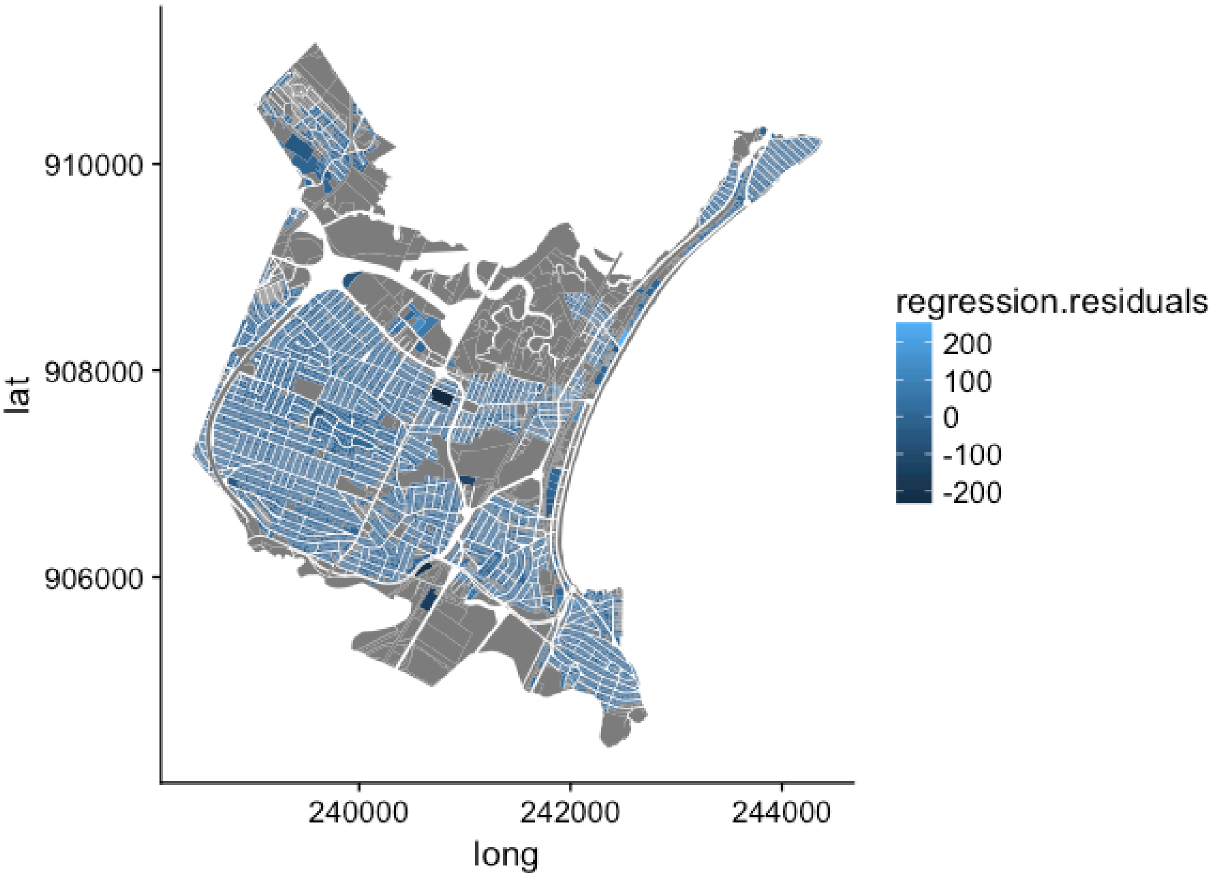
```
## OGR data source with driver: OpenFileGDB 
## Source: "/Users/vlingenfelter5/Desktop/bigData/M248_parcels_gdb/M248_parcels_sde.gdb", layer: "M2
## with 12838 features
## It has 12 fields
```

```
require(ggplot2)
require(ggmap)
require(maptools)
require(rgeos)

assessor_geo<-fortify(assessor_geo, region = "LOC_ID")
assessor_geo<-merge(assessor_geo,assessor,by.x='id',by.y='LOC_ID',duplicateGeoms=TRUE)

assessor_geo<-assessor_geo[order(assessor_geo$order),]

ggplot() +
  geom_polygon(aes(x=long, y=lat, group=group, fill=regression.residuals), data=assessor_geo)
```



We can see from this map that there is a little cluster of parcels by the beach where ther is a low number of rooms but potentially high expected value given our regression model. This suggests to me that there is a geographic element that is missing in our model.

Conclusion

From this exploratory analysis, I conclude that I want to look further into the effect the number of rooms a building has on its value. I think I would like to look at how this relationship changes based on the building code that a parcel is coded under.