Price Data Analysis

Pankaj Kumar , Karl Jurek, Shawn Jung

5/31/2019

## Introduction

Using the Modeling.csv file we created the following models:

Stepwise

LASSO

Final

to predict the cost of housing in the predictionData.csv file.

## Data Description

The modelingData.csv file contains properties sold between August 20, 2011 and June 30, 2015.

25471 observations / rows

292 variables / columns

The predictionData.csv file contains properties sold between July 2015 and May2016.

5000 observations / rows

291 variables / columns

## Data Cleaning / Wrangling

* Initial import of the the modelingData.csv and predictionData.csv required stringsAsFactors=F
* Converting Time Stamp from integer based to date format (yyyy-mm-dd)
* modeling\_df=transform(modeling\_df, timestamp=as.Date(timestamp, origin = “1899-12-30”))
* projection\_df=transform(projection\_df, timestamp=as.Date(timestamp, origin = “1899-12-30”))

### Build\_year -

* 14965 transformed to 1965
* 1691 transformed to 1991
* 20052009 transformed to 2007

### Kitch\_sq

* 2013 deleted due to being the same as the build\_year
* 2014 deleted due to being the same as the build\_year
* 1974 deleted due to being the same as the build\_year
* 1970 deleted due to being larger than the entire square footage of the building

### State

33 transformed to 3

Removed all apostrophe / single quotes from sub\_area

There are Deleted 7991 rows where the max\_floor, material, build\_year, num\_room, kitch\_sq, and state all had a value of NA.

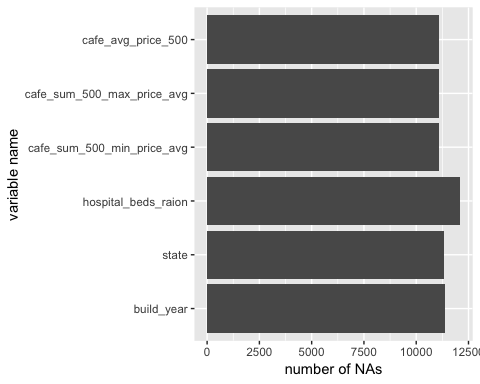
Deleted 1791 rows where the life\_sq, build\_year, and state all had a value of NA.

Deleted 1038 rows where the life\_sq and build\_year had a value of NA. Deleted 261 rows where the life\_sq and state had a value of NA. Deleted 232 rows where the build\_year and state had a value of NA.

## EDA

### Univariate Analysis

#### Number of NA’s



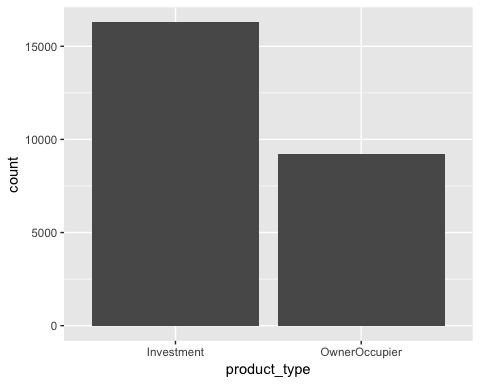
##### list of all categorical variables

##### check all factoers   
  
factors = sapply(price\_modeling\_data, function(x) class (x)=="factor")  
factor\_stack= stack(factors)  
categorical\_vars = filter(factor\_stack, values == TRUE) %>% select(ind)  
categorical\_vars

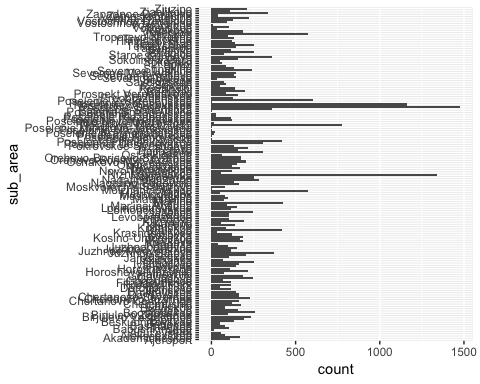
## ind  
## 1 product\_type  
## 2 sub\_area  
## 3 culture\_objects\_top\_25  
## 4 thermal\_power\_plant\_raion  
## 5 incineration\_raion  
## 6 oil\_chemistry\_raion  
## 7 radiation\_raion  
## 8 railroad\_terminal\_raion  
## 9 big\_market\_raion  
## 10 nuclear\_reactor\_raion  
## 11 detention\_facility\_raion  
## 12 water\_1line  
## 13 big\_road1\_1line  
## 14 railroad\_1line  
## 15 ecology

#### Bar plots

## we need to pass x values directly from list of categorical we got earlier   
  
ggplot(data= price\_modeling\_data)+ geom\_bar(mapping = aes(x= product\_type))



ggplot(data= price\_modeling\_data) + geom\_bar(mapping = aes(x= sub\_area))+coord\_flip()

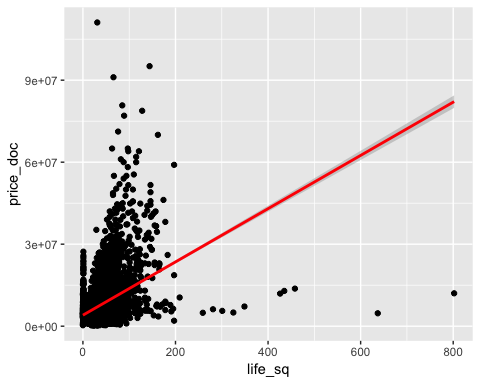


#### Bax plots

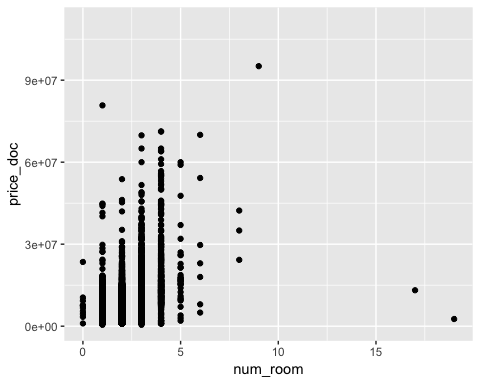
### Bivariate Analysis

## Warning: Removed 5333 rows containing non-finite values (stat\_smooth).

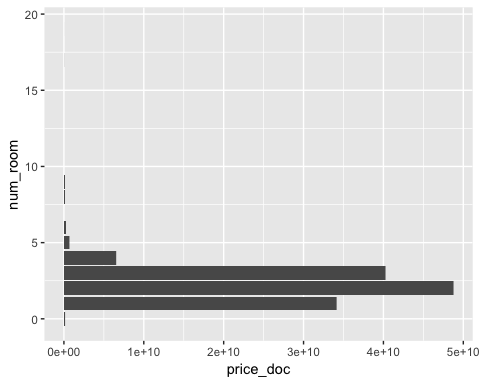
## Warning: Removed 5333 rows containing missing values (geom\_point).



## Warning: Removed 7991 rows containing missing values (geom\_point).

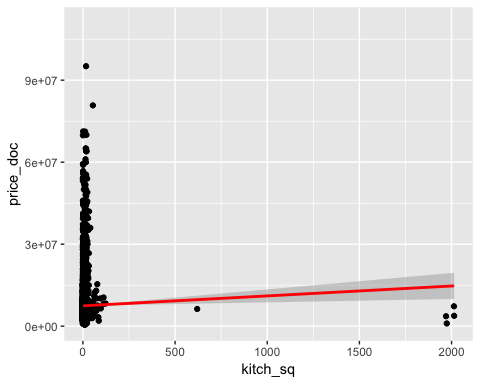


## Warning: Removed 7991 rows containing missing values (position\_stack).

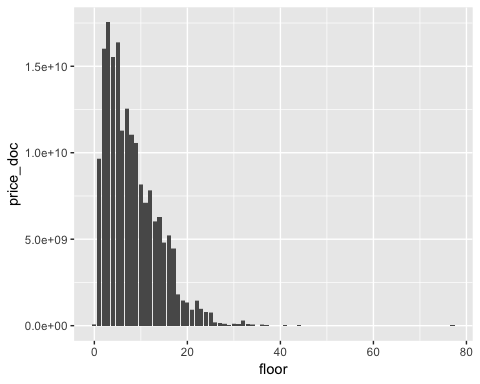


## Warning: Removed 7991 rows containing non-finite values (stat\_smooth).

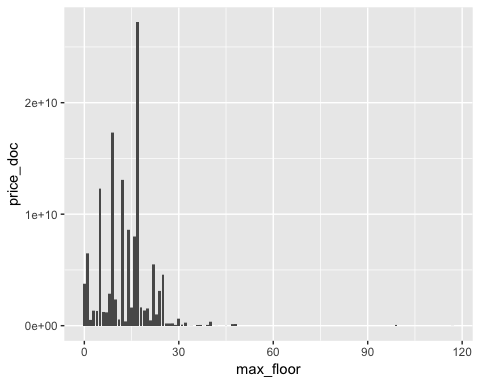
## Warning: Removed 7991 rows containing missing values (geom\_point).



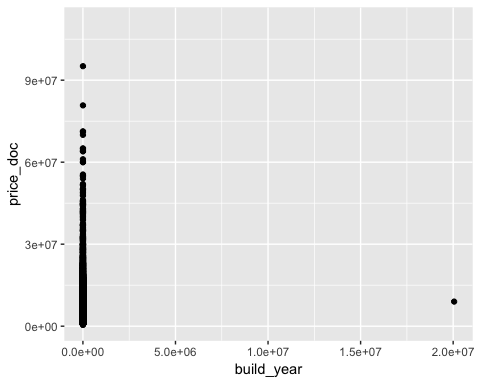
## Warning: Removed 146 rows containing missing values (position\_stack).



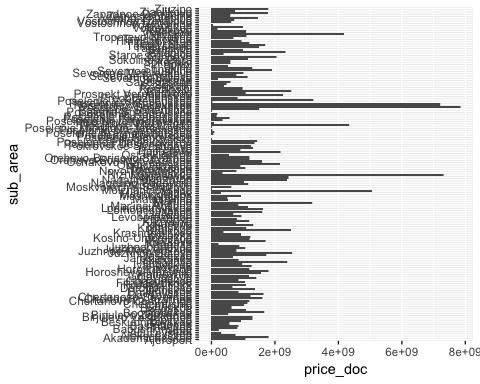
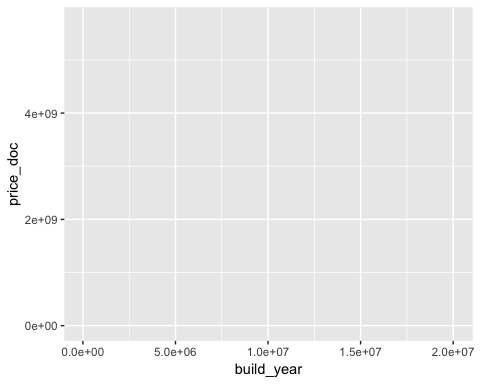
## Warning: Removed 7991 rows containing missing values (position\_stack).



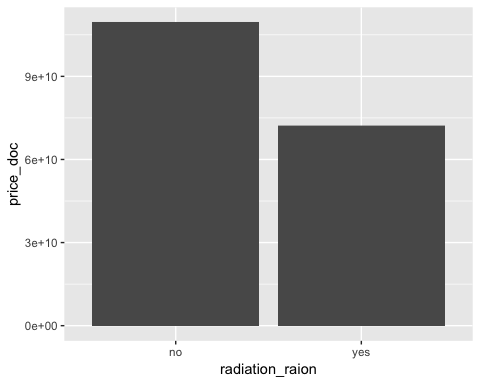
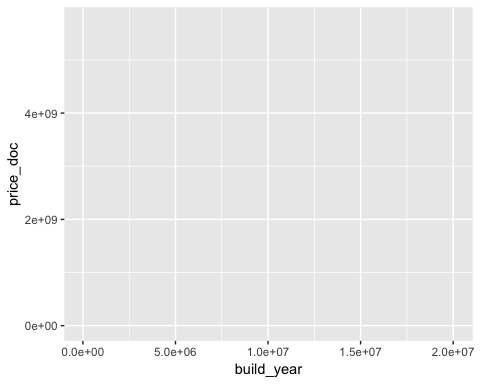
## Warning: Removed 11392 rows containing missing values (geom\_point).



## Warning: Removed 11392 rows containing missing values (position\_stack).



## Warning: Removed 11392 rows containing missing values (position\_stack).



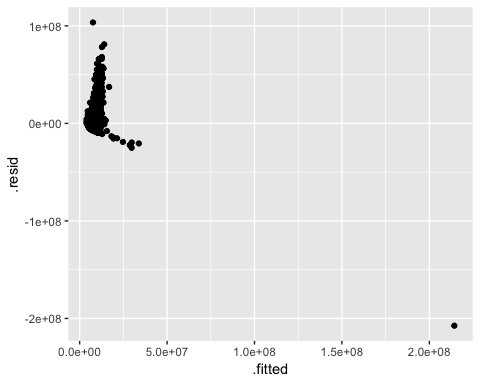
### Transformations

### Outliears

Outliers

Full\_sq - 5326

##   
## Call:  
## lm(formula = price\_doc ~ full\_sq + work\_all, data = price\_modeling\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -207412189 -1794586 -650797 929911 103496673   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.943e+06 6.312e+04 62.46 <2e-16 \*\*\*  
## full\_sq 3.915e+04 6.996e+02 55.96 <2e-16 \*\*\*  
## work\_all 1.978e+01 7.524e-01 26.29 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4511000 on 25468 degrees of freedom  
## Multiple R-squared: 0.1274, Adjusted R-squared: 0.1273   
## F-statistic: 1859 on 2 and 25468 DF, p-value: < 2.2e-16



### Interaction Terms

## Modeling

OLS  
 LASSO

## Final Prediction