Final_Project

Libaries

```
pacman::p_load(data.table,tidyverse,magrittr,YARF,skimr,plyr,tidyr,YARF,mltools,caret)
```

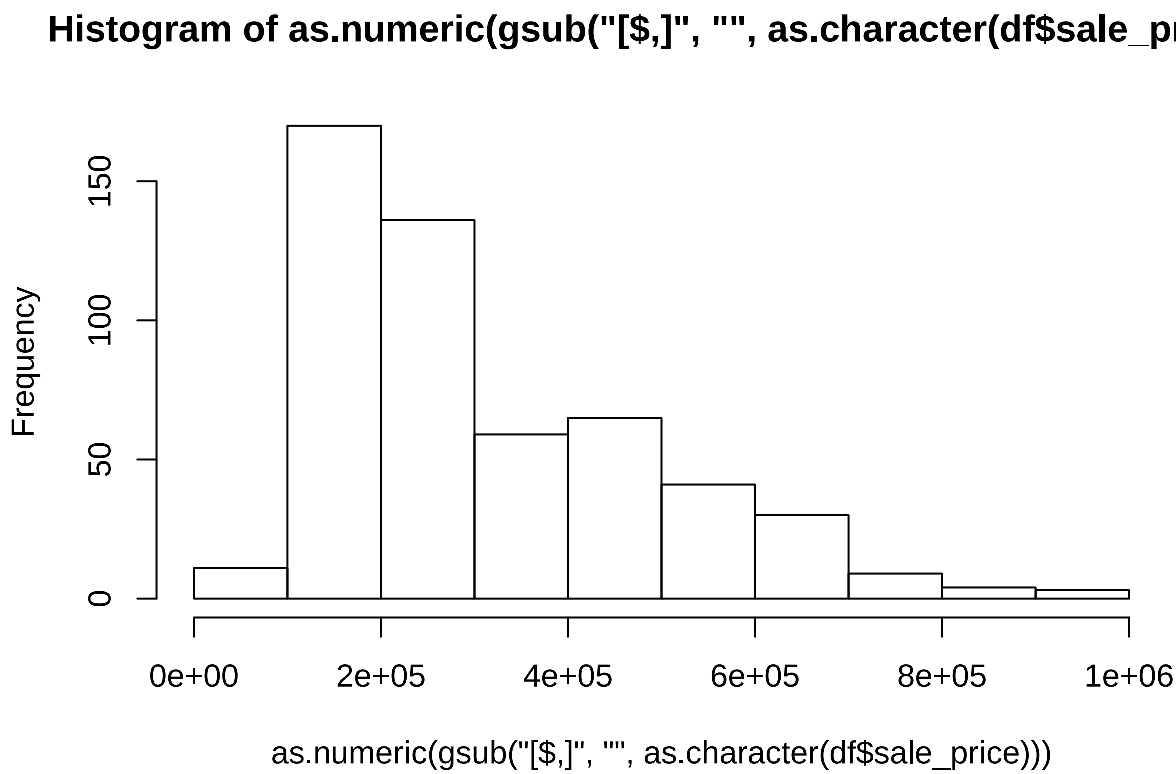
Loading Data

```
df= read.csv("housing_data_2016_2017.csv")
head(df,1)
```

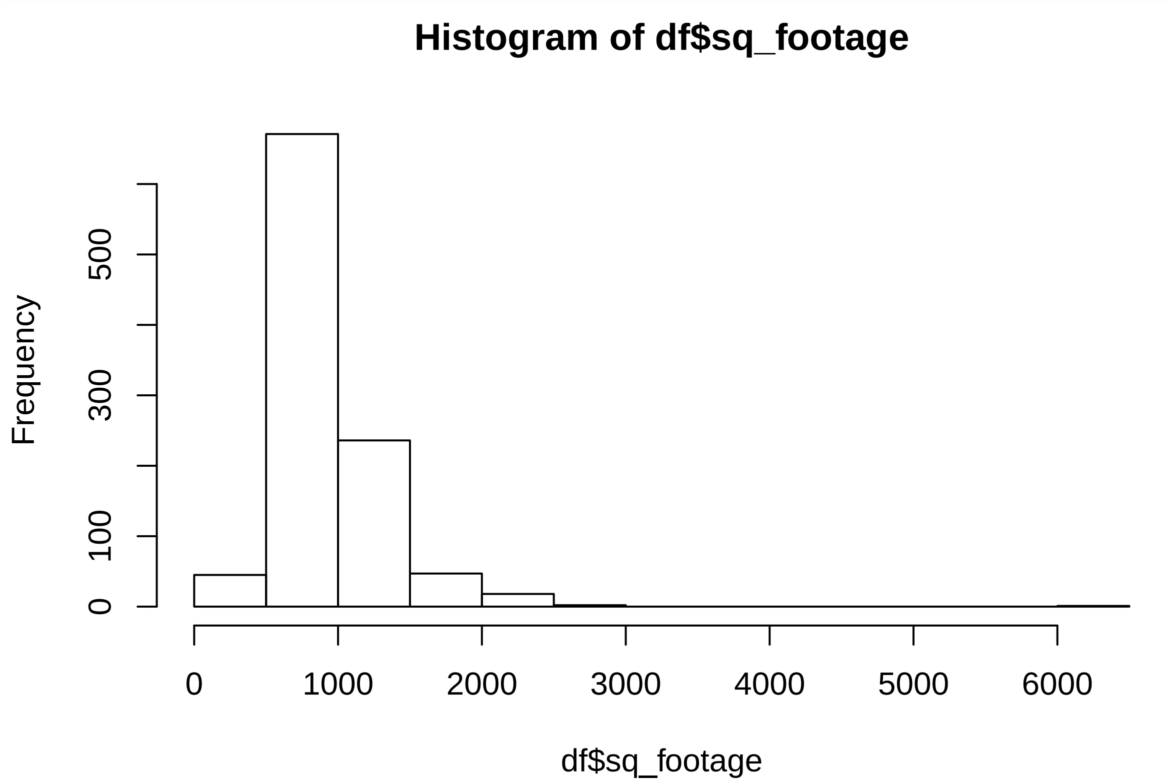
```
##                            HITId                      HITTypeId
## 1 3OID399FXG7F26JWONXF0Y86J90FD4 36BILMLQB75QQNBTYKGYCZWDN8TVAU
##                                                                Title
## 1 Find Information about Housing To Help a Student Project -- Very easy
##                                        Description Keywords Reward
## 1 Go to a link and copy information into the HIT       NA $0.05
##                   CreationTime MaxAssignments
## 1 Wed Feb 15 22:13:37 PST 2017              1
##                                  RequesterAnnotation AssignmentDurationInSeconds
## 1 BatchId:2689947;OriginalHitTemplateId:920937336;                          900
##   AutoApprovalDelayInSeconds                   Expiration NumberOfSimilarHITs
## 1                         60 Wed Feb 22 22:13:37 PST 2017                  NA
##   LifetimeInSeconds                   AssignmentId       WorkerId
## 1                NA 32KTQ2V7RDFCSAWQOW1SXC5AZIC9MB A231MNJJDDF3LS
##   AssignmentStatus                   AcceptTime                   SubmitTime
## 1         Approved Thu Feb 16 05:32:36 PST 2017 Thu Feb 16 05:35:37 PST 2017
##               AutoApprovalTime            ApprovalTime RejectionTime
## 1 Thu Feb 16 05:36:37 PST 2017 2017-02-16 13:37:11 UTC            NA
##   RequesterFeedback WorkTimeInSeconds LifetimeApprovalRate
## 1                NA               181       100% (187/187)
##   Last30DaysApprovalRate Last7DaysApprovalRate
## 1         100% (187/187)        100% (187/187)
##                                                                                               URL
## 1 http://www.mlsli.com/homes-for-sale/address-not-available-from-broker-Flushing-NY-11355-149238320
##   approx_year_built cats_allowed common_charges community_district_num
## 1              1955           no           $767                     25
##   coop_condo date_of_sale dining_room_type dogs_allowed fuel_type
## 1      co-op    2/16/2016            combo           no       gas
##   full_address_or_zip_code garage_exists kitchen_type maintenance_cost
## 1       Flushing NY, 11355          <NA>       eat in             <NA>
##          model_type num_bedrooms num_floors_in_building num_full_bathrooms
## 1 Mitchell Garden 3            2                      6                  1
##   num_half_bathrooms num_total_rooms parking_charges pct_tax_deductibl
## 1                 NA               5            <NA>                NA
##   sale_price sq_footage total_taxes walk_score listing_price_to_nearest_1000
## 1  $228,000          NA        <NA>         82                          <NA>
##    url
```

```
## 1 <NA>
```

```r
hist(as.numeric(gsub('[$,]','',as.character(df$sale_price))))
```




```r
hist(df$sq_footage)
```



##

Useful Summary of Data This gives a broad overview on the data,

```r
skim(df)
```

Table 1: Data summary

Name	df
Number of rows	2230
Number of columns	55
Column type frequency:	
factor	36
logical	5
numeric	14
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
HITId	758	0.66	FALSE	1472	301: 1, 301: 1, 301: 1, 302: 1
HITTypeId	758	0.66	FALSE	2	310: 944, 36B: 528
Title	758	0.66	FALSE	1	Fin: 1472
Description	758	0.66	FALSE	2	Got: 944, Go : 528
Reward	758	0.66	FALSE	1	$0.: 1472
CreationTime	758	0.66	FALSE	62	Thu: 43, Thu: 40, Wed: 39, Thu: 37
RequesterAnnotation	758	0.66	FALSE	2	Bat: 944, Bat: 528
Expiration	758	0.66	FALSE	62	Thu: 43, Thu: 40, Wed: 39, Thu: 37
AssignmentId	758	0.66	FALSE	1472	301: 1, 301: 1, 304: 1, 304: 1
WorkerId	758	0.66	FALSE	73	A23: 187, A1S: 129, A3C: 124, AHX
AssignmentStatus	758	0.66	FALSE	1	App: 1472
AcceptTime	758	0.66	FALSE	1457	Thu: 2, Thu: 2, Thu: 2, Thu: 2
SubmitTime	758	0.66	FALSE	1460	Thu: 2, Thu: 2, Thu: 2, Thu: 2
AutoApprovalTime	758	0.66	FALSE	1460	Thu: 2, Thu: 2, Thu: 2, Thu: 2
ApprovalTime	758	0.66	FALSE	929	201: 6, 201: 6, 201: 5, 201: 5
LifetimeApprovalRate	758	0.66	FALSE	32	100: 187, 100: 126, 100: 124, 100: 10
Last30DaysApprovalRate	758	0.66	FALSE	32	100: 187, 100: 126, 100: 124, 100: 10
Last7DaysApprovalRate	758	0.66	FALSE	32	100: 187, 100: 126, 100: 124, 100: 10
URL	758	0.66	FALSE	1450	htt: 2, htt: 2, htt: 2, htt: 2
cats_allowed	0	1.00	FALSE	3	no: 1402, yes: 826, y: 2
common_charges	1684	0.24	FALSE	258	$25: 11, $17: 10, $27: 9, $29: 8
coop_condo	0	1.00	FALSE	2	co-: 1661, con: 569
date_of_sale	1702	0.24	FALSE	222	6/3: 7, 10/: 6, 12/: 6, 2/2: 6
dining_room_type	448	0.80	FALSE	5	com: 957, for: 620, oth: 201, din: 2
dogs_allowed	0	1.00	FALSE	3	no: 1684, yes: 544, yes: 2
fuel_type	112	0.95	FALSE	6	gas: 1348, oil: 664, ele: 62, oth: 40
full_address_or_zip_code	0	1.00	FALSE	1177	70-: 22, 269: 17, 270: 16, 73-: 14
garage_exists	1826	0.18	FALSE	6	yes: 361, Yes: 39, 1: 1, eys: 1
kitchen_type	16	0.99	FALSE	13	eat: 733, eff: 505, com: 349, eff: 338
maintenance_cost	623	0.72	FALSE	609	$54: 10, $67: 10, $68: 10, $70: 10
model_type	40	0.98	FALSE	875	1 B: 63, One: 59, 2 B: 50, Hi-: 41
parking_charges	1671	0.25	FALSE	89	$15: 42, $60: 41, $75: 27, $13: 23
sale_price	1702	0.24	FALSE	315	$15: 11, $17: 10, $13: 7, $22: 7

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
total_taxes	1646	0.26	FALSE	293	$13: 13, $25: 12, $4,: 11, $2,: 10
listing_price_to_nearest_1000	534	0.76	FALSE	292	$34: 28, $39: 26, $28: 25, $23: 23
url	758	0.66	FALSE	1450	htt: 2, htt: 2, htt: 2, htt: 2

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
Keywords	2230	0	NaN	:
NumberOfSimilarHITs	2230	0	NaN	:
LifetimeInSeconds	2230	0	NaN	:
RejectionTime	2230	0	NaN	:
RequesterFeedback	2230	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
MaxAssignments	758	0.66	1.00	0.00	1	1	1	1	1	
AssignmentDurationInSeconds	758	0.66	900.00	0.00	900	900	900	900	900	
AutoApprovalDelayInSeconds	758	0.66	60.00	0.00	60	60	60	60	60	
WorkTimeInSeconds	758	0.66	162.39	111.69	22	89	127	197	815	
approx_year_built	40	0.98	1962.71	21.08	1893	1950	1958	1970	2017	
community_district_num	19	0.99	26.33	2.95	3	25	26	28	32	
num_bedrooms	115	0.95	1.65	0.74	0	1	2	2	6	
num_floors_in_building	650	0.71	7.79	7.52	1	3	6	7	34	
num_full_bathrooms	0	1.00	1.23	0.44	1	1	1	1	3	
num_half_bathrooms	2058	0.08	0.95	0.30	0	1	1	1	2	
num_total_rooms	2	1.00	4.14	1.35	0	3	4	5	14	
pct_tax_deductibl	1754	0.21	45.40	6.95	20	40	50	50	75	
sq_footage	1210	0.46	955.36	380.86	100	743	881	1100	6215	
walk_score	0	1.00	83.92	14.75	7	77	89	95	99	

There is a lot of data that is completely missing and some that is heavily missing. I decided to remove them. Some examples below.

Keywords,NumberOfSimilarHITs, LigetimeInSeconds, RejectionTime,RequesterFeedback all completely missing. ommon_charges(missing 1684),garage_exists(missing 1826)

```
cat("Data has",nrow(df),"number of rows\n")
```

```
## Data has 2230 number of rows
```

```
cat("Data has",ncol(df), "number of columns")
```

```
## Data has 55 number of columns
```

```
sort(colMeans(is.na(df)), decreasing = TRUE)
```

```
##                  Keywords         NumberOfSimilarHITs
##               1.000000000                 1.000000000
##         LifetimeInSeconds               RejectionTime
```

```
##                        1.000000000                 1.000000000
##                  RequesterFeedback           num_half_bathrooms
##                        1.000000000                 0.922869955
##                      garage_exists           pct_tax_deductibl
##                        0.818834081                 0.786547085
##                       date_of_sale                  sale_price
##                        0.763228700                 0.763228700
##                     common_charges             parking_charges
##                        0.755156951                 0.749327354
##                        total_taxes                         url
##                        0.738116592                 0.660089686
##                         sq_footage                       HITId
##                        0.542600897                 0.339910314
##                          HITTypeId                       Title
##                        0.339910314                 0.339910314
##                        Description                      Reward
##                        0.339910314                 0.339910314
##                       CreationTime              MaxAssignments
##                        0.339910314                 0.339910314
##                RequesterAnnotation AssignmentDurationInSeconds
##                        0.339910314                 0.339910314
##          AutoApprovalDelayInSeconds                 Expiration
##                        0.339910314                 0.339910314
##                       AssignmentId                    WorkerId
##                        0.339910314                 0.339910314
##                   AssignmentStatus                  AcceptTime
##                        0.339910314                 0.339910314
##                         SubmitTime            AutoApprovalTime
##                        0.339910314                 0.339910314
##                       ApprovalTime            WorkTimeInSeconds
##                        0.339910314                 0.339910314
##               LifetimeApprovalRate      Last30DaysApprovalRate
##                        0.339910314                 0.339910314
##                Last7DaysApprovalRate                        URL
##                        0.339910314                 0.339910314
##             num_floors_in_building            maintenance_cost
##                        0.291479821                 0.279372197
## listing_price_to_nearest_1000             dining_room_type
##                        0.239461883                 0.200896861
##                       num_bedrooms                   fuel_type
##                        0.051569507                 0.050224215
##                   approx_year_built                  model_type
##                        0.017937220                 0.017937220
##             community_district_num                kitchen_type
##                        0.008520179                 0.007174888
##                    num_total_rooms                cats_allowed
##                        0.000896861                 0.000000000
##                         coop_condo                dogs_allowed
##                        0.000000000                 0.000000000
##            full_address_or_zip_code          num_full_bathrooms
##                        0.000000000                 0.000000000
##                         walk_score
##                        0.000000000
```

Data Cleaning Remove all missing y

```
df_drops = df %>% drop_na(sale_price)
# skim(df_drops) %>%
#   summary()
skim(df_drops)
```

Table 5: Data summary

Name	df_drops
Number of rows	528
Number of columns	55
Column type frequency:	
factor	36
logical	5
numeric	14
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
HITId	0	1.00	FALSE	528	301: 1, 302: 1, 302: 1, 307: 1
HITTypeId	0	1.00	FALSE	1	36B: 528, 310: 0
Title	0	1.00	FALSE	1	Fin: 528
Description	0	1.00	FALSE	1	Go : 528, Got: 0
Reward	0	1.00	FALSE	1	$0.: 528
CreationTime	0	1.00	FALSE	21	Wed: 39, Wed: 36, Wed: 33, Wed: 3
RequesterAnnotation	0	1.00	FALSE	1	Bat: 528, Bat: 0
Expiration	0	1.00	FALSE	21	Wed: 39, Wed: 36, Wed: 33, Wed: 3
AssignmentId	0	1.00	FALSE	528	301: 1, 301: 1, 308: 1, 308: 1
WorkerId	0	1.00	FALSE	21	A23: 187, AHX: 102, A1K: 80, A3S:
AssignmentStatus	0	1.00	FALSE	1	App: 528
AcceptTime	0	1.00	FALSE	523	Thu: 2, Thu: 2, Thu: 2, Thu: 2
SubmitTime	0	1.00	FALSE	524	Thu: 2, Thu: 2, Thu: 2, Thu: 2
AutoApprovalTime	0	1.00	FALSE	524	Thu: 2, Thu: 2, Thu: 2, Thu: 2
ApprovalTime	0	1.00	FALSE	337	201: 5, 201: 5, 201: 4, 201: 4
LifetimeApprovalRate	0	1.00	FALSE	15	100: 187, 100: 102, 100: 80, 100: 58
Last30DaysApprovalRate	0	1.00	FALSE	15	100: 187, 100: 102, 100: 80, 100: 58
Last7DaysApprovalRate	0	1.00	FALSE	15	100: 187, 100: 102, 100: 80, 100: 58
URL	0	1.00	FALSE	524	htt: 2, htt: 2, htt: 2, htt: 2
cats_allowed	0	1.00	FALSE	2	no: 285, yes: 243, y: 0
common_charges	396	0.25	FALSE	112	$27: 3, $31: 3, $21: 2, $24: 2
coop_condo	0	1.00	FALSE	2	co-: 399, con: 129
date_of_sale	0	1.00	FALSE	222	6/3: 7, 10/: 6, 12/: 6, 2/2: 6
dining_room_type	120	0.77	FALSE	4	com: 241, for: 116, oth: 49, din: 2
dogs_allowed	0	1.00	FALSE	2	no: 381, yes: 147, yes: 0
fuel_type	24	0.95	FALSE	6	gas: 301, oil: 180, ele: 11, oth: 8
full_address_or_zip_code	0	1.00	FALSE	468	70-: 8, 54-: 4, 104: 3, 117: 3
garage_exists	434	0.18	FALSE	6	yes: 51, Yes: 39, 1: 1, eys: 1
kitchen_type	6	0.99	FALSE	7	eff: 231, eat: 190, Com: 50, com: 31

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
maintenance_cost	142	0.73	FALSE	284	$52: 4, $60: 4, $66: 4, $67: 4
model_type	15	0.97	FALSE	356	1 B: 23, One: 19, 2 B: 11, Gar: 11
parking_charges	393	0.26	FALSE	50	$10: 12, $20: 10, $95: 8, $12: 7
sale_price	0	1.00	FALSE	315	$15: 11, $17: 10, $13: 7, $22: 7
total_taxes	397	0.25	FALSE	120	$2,: 3, $4,: 3, $1,: 2, $1,: 2
listing_price_to_nearest_1000	528	0.00	FALSE	0	$1,: 0, $10: 0, $10: 0, $10: 0
url	0	1.00	FALSE	524	htt: 2, htt: 2, htt: 2, htt: 2

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
Keywords	528	0	NaN	:
NumberOfSimilarHITs	528	0	NaN	:
LifetimeInSeconds	528	0	NaN	:
RejectionTime	528	0	NaN	:
RequesterFeedback	528	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
MaxAssignments	0	1.00	1.00	0.00	1	1	1	1	1	
AssignmentDurationInSeconds	0	1.00	900.00	0.00	900	900	900	900	900	
AutoApprovalDelayInSeconds	0	1.00	60.00	0.00	60	60	60	60	60	
WorkTimeInSeconds	0	1.00	150.38	96.99	52	97	123	163	815	
approx_year_built	6	0.99	1962.38	20.56	1915	1950	1957	1968	2016	
community_district_num	1	1.00	26.30	2.99	3	25	26	28	30	
num_bedrooms	0	1.00	1.54	0.75	0	1	1	2	3	
num_floors_in_building	108	0.80	7.08	6.83	1	2	6	7	34	
num_full_bathrooms	0	1.00	1.20	0.42	1	1	1	1	3	
num_half_bathrooms	498	0.06	1.03	0.18	1	1	1	1	2	
num_total_rooms	0	1.00	4.02	1.20	1	3	4	5	8	
pct_tax_deductibl	429	0.19	44.99	8.09	20	40	50	50	65	
sq_footage	315	0.40	965.28	490.42	375	750	874	1010	6215	
walk_score	0	1.00	83.10	13.09	15	76	85	94	99	

Meaningful Features Data Cleaning

Finding meaningful features. These are features I believe are meaningful. df_mutated has all the features that I will be using. I am not looking at whats missing yet or how the data looks like, just looking for features that would be best to predict sales price.

```
colnames(df)
```

```
##  [1] "HITId"                       "HITTypeId"
##  [3] "Title"                       "Description"
##  [5] "Keywords"                    "Reward"
##  [7] "CreationTime"                "MaxAssignments"
##  [9] "RequesterAnnotation"         "AssignmentDurationInSeconds"
## [11] "AutoApprovalDelayInSeconds"  "Expiration"
```

```
## [13] "NumberOfSimilarHITs"           "LifetimeInSeconds"
## [15] "AssignmentId"                  "WorkerId"
## [17] "AssignmentStatus"              "AcceptTime"
## [19] "SubmitTime"                    "AutoApprovalTime"
## [21] "ApprovalTime"                  "RejectionTime"
## [23] "RequesterFeedback"             "WorkTimeInSeconds"
## [25] "LifetimeApprovalRate"          "Last30DaysApprovalRate"
## [27] "Last7DaysApprovalRate"         "URL"
## [29] "approx_year_built"             "cats_allowed"
## [31] "common_charges"                "community_district_num"
## [33] "coop_condo"                    "date_of_sale"
## [35] "dining_room_type"              "dogs_allowed"
## [37] "fuel_type"                     "full_address_or_zip_code"
## [39] "garage_exists"                 "kitchen_type"
## [41] "maintenance_cost"              "model_type"
## [43] "num_bedrooms"                  "num_floors_in_building"
## [45] "num_full_bathrooms"            "num_half_bathrooms"
## [47] "num_total_rooms"               "parking_charges"
## [49] "pct_tax_deductibl"             "sale_price"
## [51] "sq_footage"                    "total_taxes"
## [53] "walk_score"                    "listing_price_to_nearest_1000"
## [55] "url"
```

```
df_mutated = copy(df_drops)
df_mutated %<>%
  select(cats_allowed,common_charges,coop_condo,dining_room_type,dogs_allowed,fuel_type,garage_exists,ma
sort(colMeans(is.na(df_mutated)), decreasing = TRUE)
```

```
##          garage_exists            total_taxes         common_charges
##            0.821969697            0.751893939            0.750000000
##             sq_footage       maintenance_cost       dining_room_type
##            0.596590909            0.268939394            0.227272727
## num_floors_in_building              fuel_type             model_type
##            0.204545455            0.045454545            0.028409091
##      approx_year_built community_district_num           cats_allowed
##            0.011363636            0.001893939            0.000000000
##             coop_condo           dogs_allowed           num_bedrooms
##            0.000000000            0.000000000            0.000000000
##     num_full_bathrooms        num_total_rooms             sale_price
##            0.000000000            0.000000000            0.000000000
##             walk_score
##            0.000000000
```

Feature Data Cleaning

I am now looking more closely to the data.Looking at this there are too many types of model_types 875 different times from original data with NA sale price this seems difficult to deal with so I will remove this. I discarded data with more than 50% of missinginess.

```
df_mutated_features = copy(df_mutated)
df_mutated_features %<>%
  select(-model_type,-total_taxes, -community_district_num)#,-common_charges,-sq_footage)
skim(df_mutated_features) %>%
  summary()
```

Table 9: Data summary

Name	df_mutated_features
Number of rows	528
Number of columns	16
Column type frequency:	
factor	9
numeric	7
Group variables	None

```
sort(colMeans(is.na(df_mutated_features)), decreasing = TRUE)
```

```
##         garage_exists        common_charges            sq_footage
##            0.82196970            0.75000000            0.59659091
##      maintenance_cost      dining_room_type num_floors_in_building
##            0.26893939            0.22727273            0.20454545
##             fuel_type      approx_year_built          cats_allowed
##            0.04545455            0.01136364            0.00000000
##            coop_condo          dogs_allowed          num_bedrooms
##            0.00000000            0.00000000            0.00000000
##    num_full_bathrooms       num_total_rooms            sale_price
##            0.00000000            0.00000000            0.00000000
##            walk_score
##            0.00000000
```

Oberservations Data Cleaning

I am okay with the number of features I have now. Now Ill be cleaning the observations.

```
df_clean = copy(df_mutated_features)


# Fixing y to be just yes and reducing factors to just yes and no.
df_clean %<>%
  mutate(cats_allowed = as.factor(ifelse(cats_allowed =='y' | cats_allowed =='yes','yes','no'))) %>%

#Fixing yes89 to just yes and reducing factors to just yes and no
  mutate(dogs_allowed = as.factor(ifelse(dogs_allowed =='yes89' | dogs_allowed =='yes','yes','no'))) %>

  #mutate(sale_price = as.numeric(gsub('[$]','',as.character(df_clean$sale_price))))
  mutate(sale_price = as.numeric(gsub('[$,]','',as.character(df_clean$sale_price))) )%>%

  mutate(common_charges = as.numeric(gsub('[$,]','',as.character(df_clean$common_charges)))) %>%

  mutate(maintenance_cost = as.numeric(gsub('[$,]','',as.character(df_clean$maintenance_cost)))) %>%

  mutate(garage_exists = ifelse(is.na(garage_exists), 0, 1))

  #mutate(fuel_type = if(is.na(fuel_type)){fuel_type = 'other'})
```

```
  #Very annoying this best way I found to combine two factor lvels


library(forcats)
df_clean$fuel_type = fct_collapse(df_clean$fuel_type, other = c("other","Other"))
#df_clean$dining_room_type = fct_collapse(df_clean$dining_room_type, other= c('other','none','dining ar




# df_clean_sub = copy(df_clean)
# df_clean_sub = df_clean_sub[df_clean_sub$sale_price < 700000,]
#
# df_clean = df_clean_sub
 #df_clean = df_clean[df_clean$sq_footage < 2500,]
#which(df_clean$sq_footage > 2500)

#df_clean = df_clean[-136,]
```

```
options(scipen=999)
max(df_clean$sq_footage, na.rm = TRUE)
```

```
## [1] 6215
```

```
min(df_clean$sq_footage, na.rm = TRUE)
```

```
## [1] 375
```

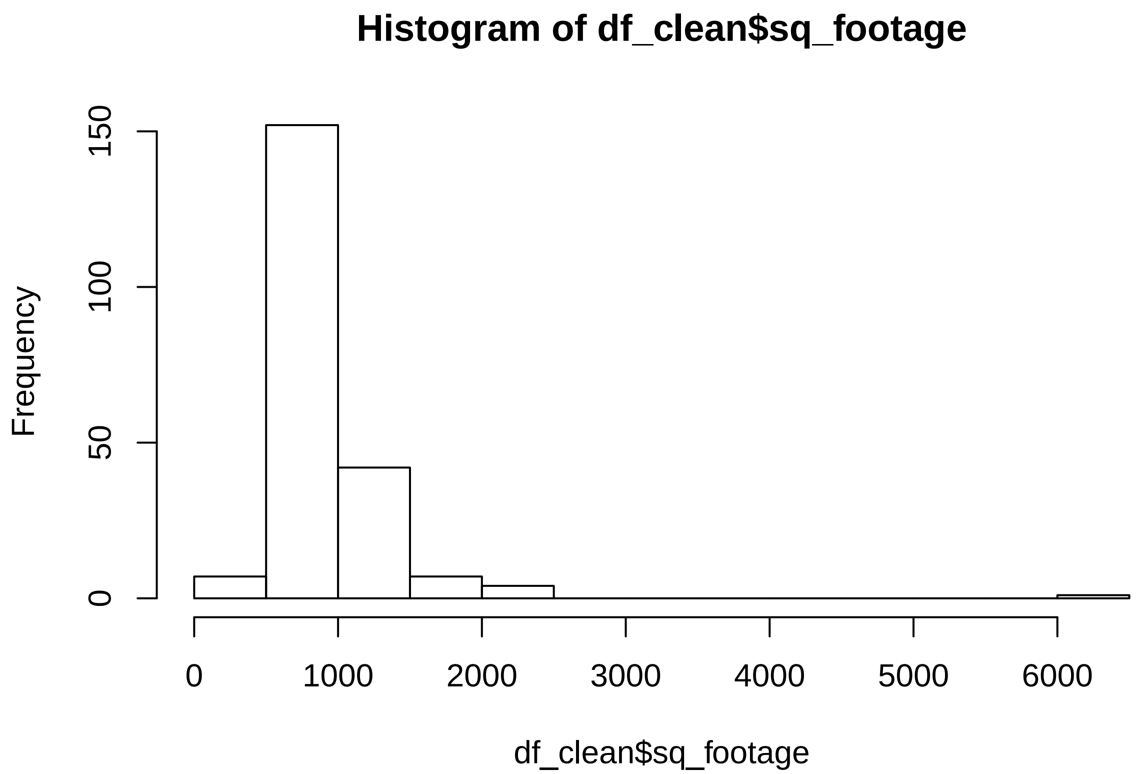
```
max(df_clean$sale_price, na.rm = TRUE)
```

```
## [1] 999999
```

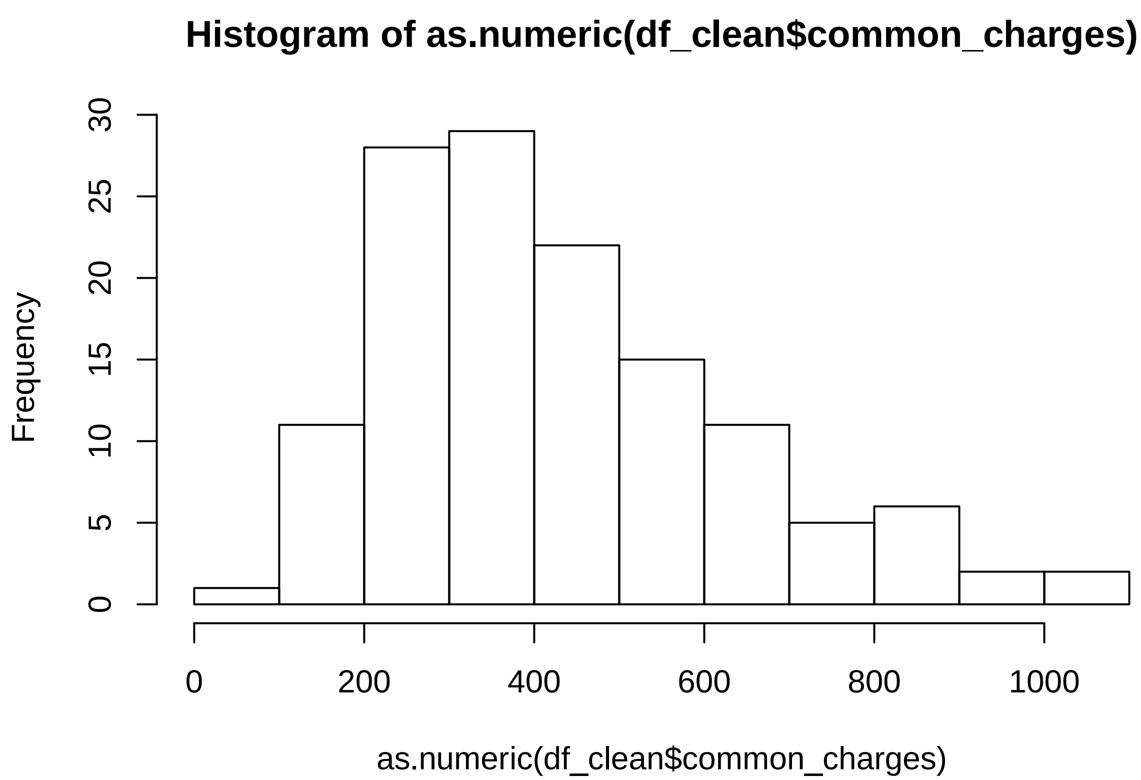
```
min(df_clean$sale_price, na.rm = TRUE)
```

```
## [1] 55000
```

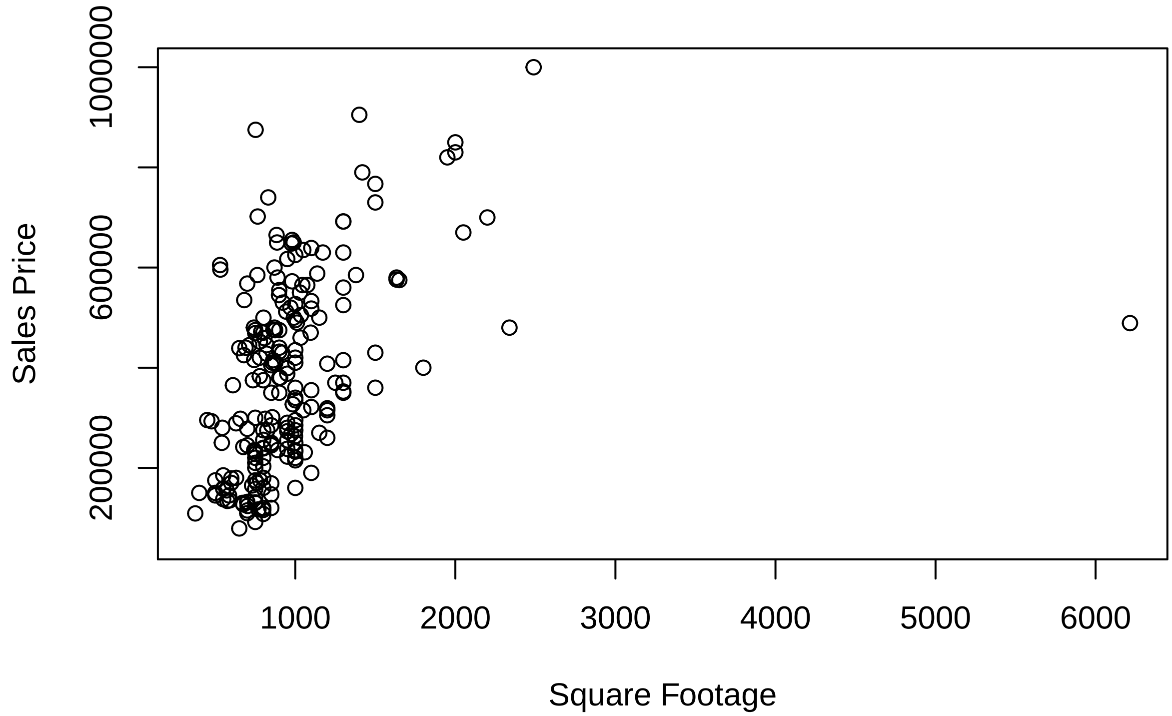
```
hist(df_clean$sq_footage)
```




```
hist(as.numeric(df_clean$common_charges))
```




```
plot(y=df_clean$sale_price,df_clean$sq_footage, xlab ='Square Footage', ylab= 'Sales Price',)
```




```
library(ggplot2)
#plot(y=df_final$sale_price,df_clean$approx_year_built, xlab ='Approx Year Built', ylab= 'Sales Price',
#geom_rug(sides ="bl")
```

```
#df_clean %<>%
  #select(-sq_footage)
```

```
#Fix issue when using MLR, Notice that there is missingness so ill set it to other
df_clean$fuel_type[df_clean$fuel_type == 'none'] = 'other'
df_clean$fuel_type[is.na(df_clean$fuel_type)] = 'other'
df_clean$fuel_type = factor(df_clean$fuel_type)
#df_cleandd = df_clean[df_clean$sq_footage < 2500,]
#df_clean = subset(df_clean, df_clean$sq_footage <= 2500 & !is.na(df_clean$sq_footage))
#df_clean$num_full_bathrooms = as.factor(df_clean$num_full_bathrooms)
skim(df_clean)
```

Table 10: Data summary

Name	df_clean
Number of rows	528
Number of columns	16
Column type frequency:	
factor	5
numeric	11
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cats_allowed	0	1.00	FALSE	2	no: 285, yes: 243
coop_condo	0	1.00	FALSE	2	co-: 399, con: 129
dining_room_type	120	0.77	FALSE	4	com: 241, for: 116, oth: 49, din: 2
dogs_allowed	0	1.00	FALSE	2	no: 381, yes: 147
fuel_type	0	1.00	FALSE	4	gas: 301, oil: 180, oth: 36, ele: 11

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p7
common_charges	396	0.25	433.92	205.40	70	288.50	390.5	537.7
garage_exists	0	1.00	0.18	0.38	0	0.00	0.0	0.0
maintenance_cost	142	0.73	821.85	378.77	155	639.25	734.0	880.0
approx_year_built	6	0.99	1962.38	20.56	1915	1950.00	1957.0	1968.0
num_bedrooms	0	1.00	1.54	0.75	0	1.00	1.0	2.0
num_floors_in_building	108	0.80	7.08	6.83	1	2.00	6.0	7.0
num_full_bathrooms	0	1.00	1.20	0.42	1	1.00	1.0	1.0
num_total_rooms	0	1.00	4.02	1.20	1	3.00	4.0	5.0
sq_footage	315	0.40	965.28	490.42	375	750.00	874.0	1010.0
sale_price	0	1.00	314956.56	179526.60	55000	171500.00	259500.0	428875.0
walk_score	0	1.00	83.10	13.09	15	76.00	85.0	94.0

```r
set.seed(28)
M = tbl_df(apply(is.na(df_clean), 2, as.numeric))
colnames(M) = paste("is_missing_", colnames(df_clean), sep = "")
M %<>%
  select_if(function(x){sum(x) > 0})
head(M)
```

```
## # A tibble: 6 x 6
##   is_missing_comm~ is_missing_dini~ is_missing_main~ is_missing_appr~
##              <dbl>            <dbl>            <dbl>            <dbl>
## 1                0                0                1                0
## 2                1                0                0                0
## 3                0                0                1                0
## 4                0                0                1                0
## 5                1                0                0                0
## 6                1                0                0                0
## # ... with 2 more variables: is_missing_num_floors_in_building <dbl>,
## #   is_missing_sq_footage <dbl>
```

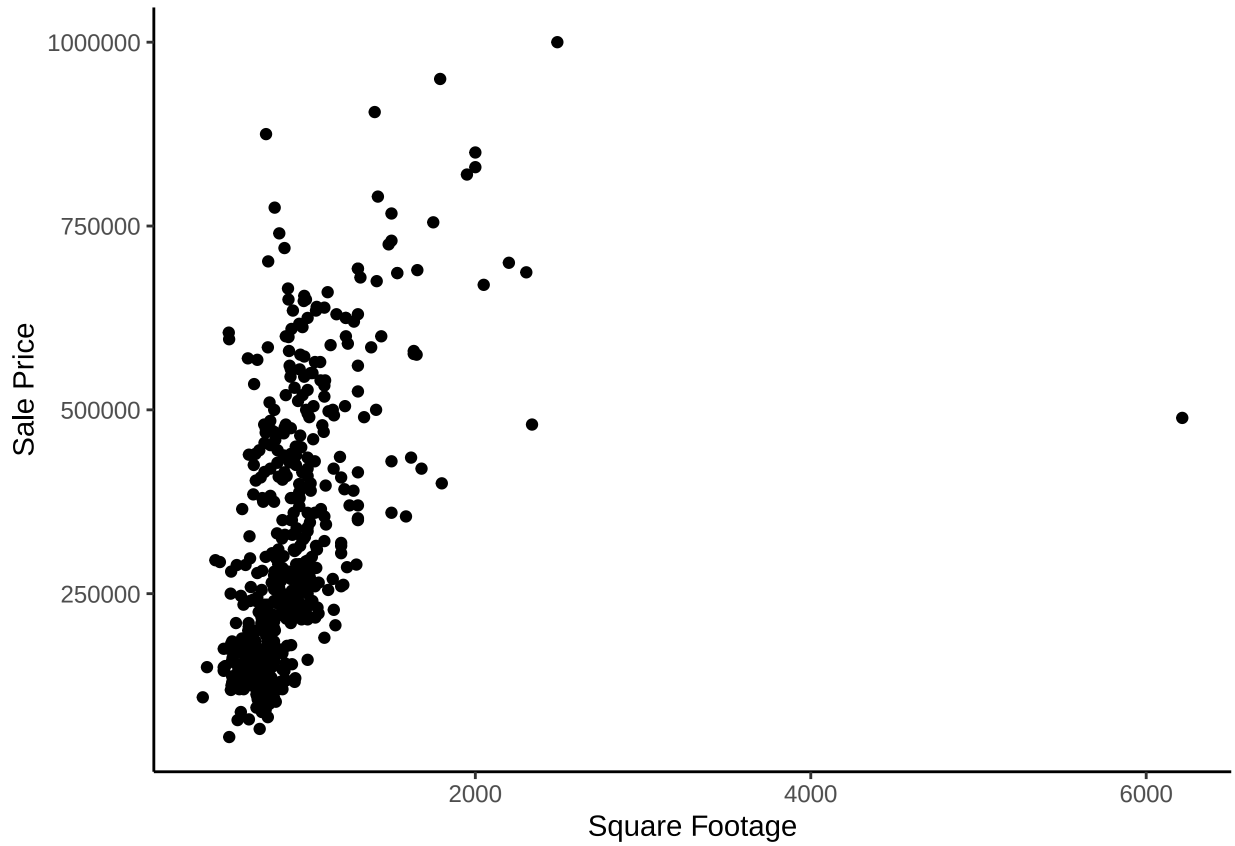
```r
pacman::p_load(missForest)
dfimp = missForest(data.frame(df_clean))$ximp
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
##   missForest iteration 6 in progress...done!
##   missForest iteration 7 in progress...done!
```

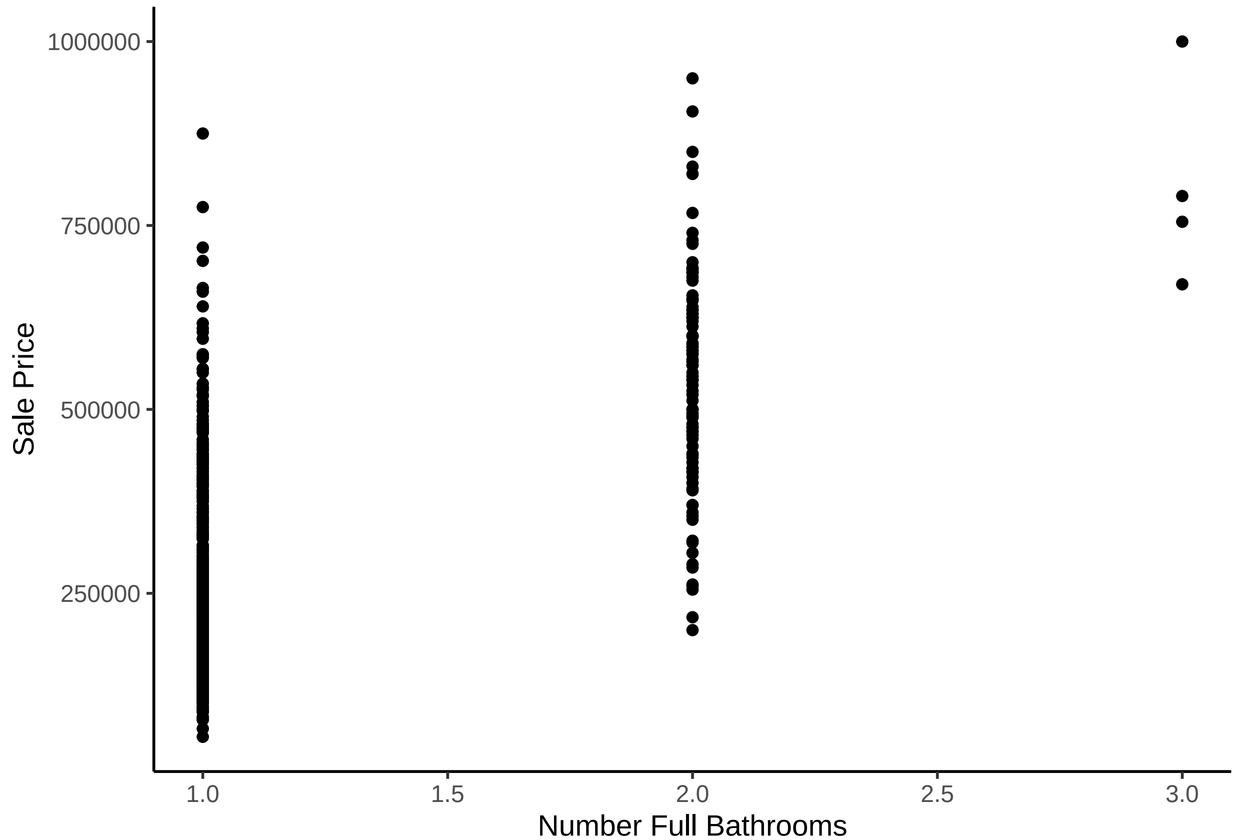
```r
df_final = cbind(dfimp, M)
```

```r
#skim(df_final)
```

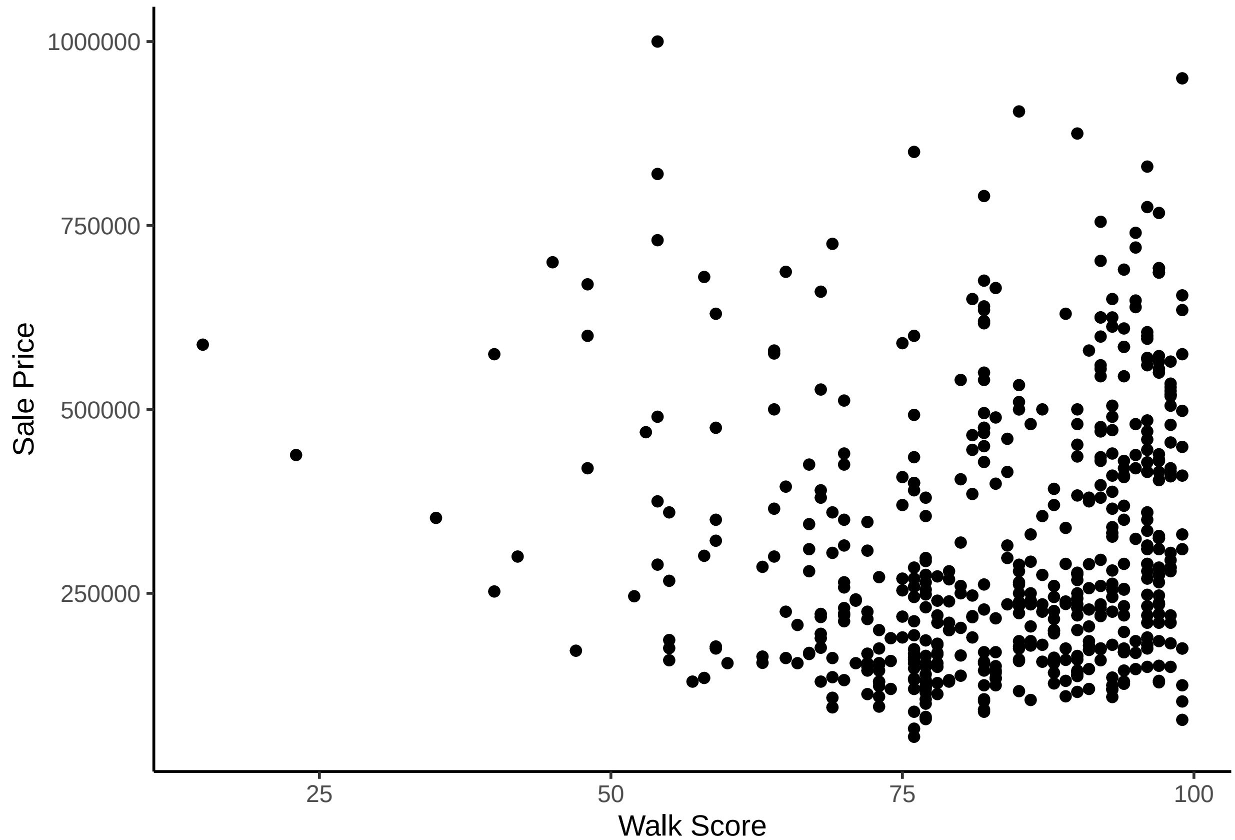
```r
ggplot(df_final, aes(x=sq_footage, y=sale_price)) + geom_point() +  theme_classic()  +
  labs(
       x="Square Footage", y = "Sale Price")
```



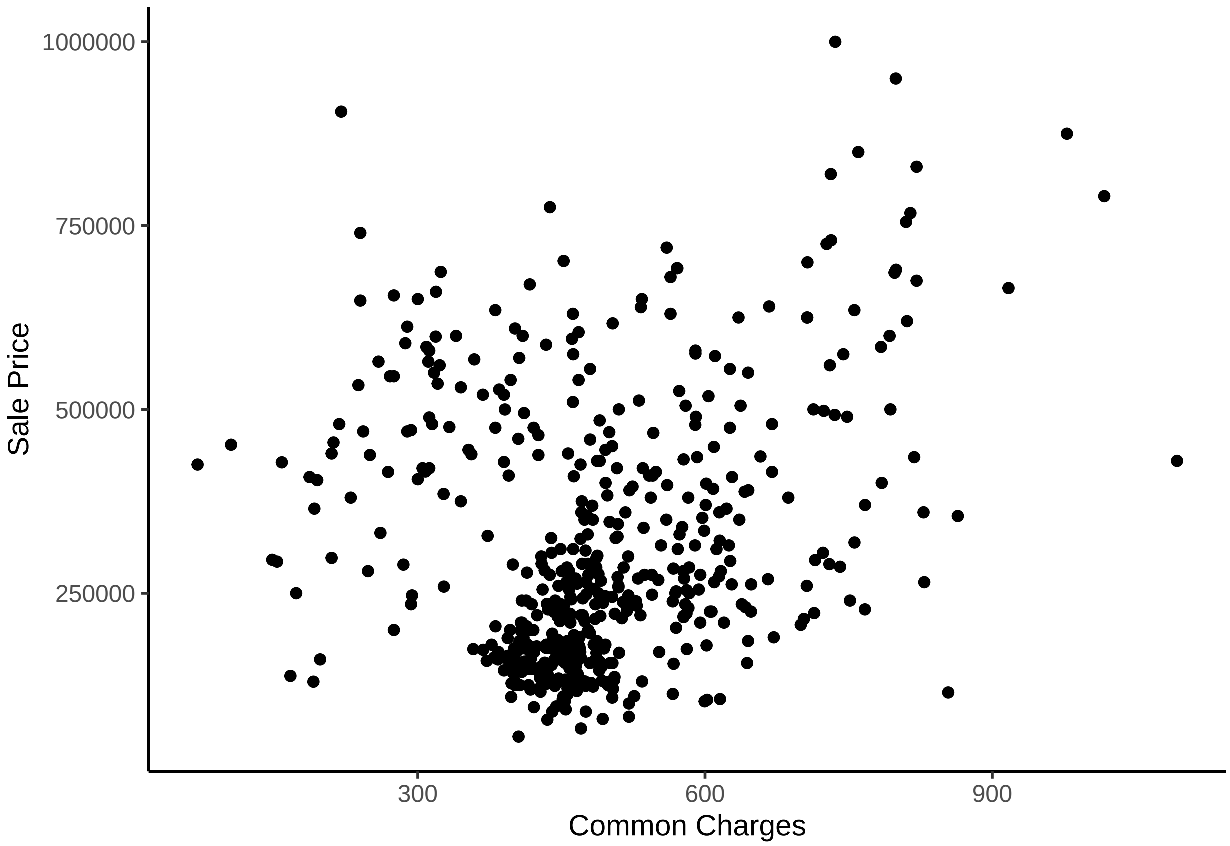

```r
ggplot(df_final, aes(x=num_full_bathrooms, y=sale_price)) + geom_point() +  theme_classic()  +
  labs(
       x="Number Full Bathrooms", y = "Sale Price")
```



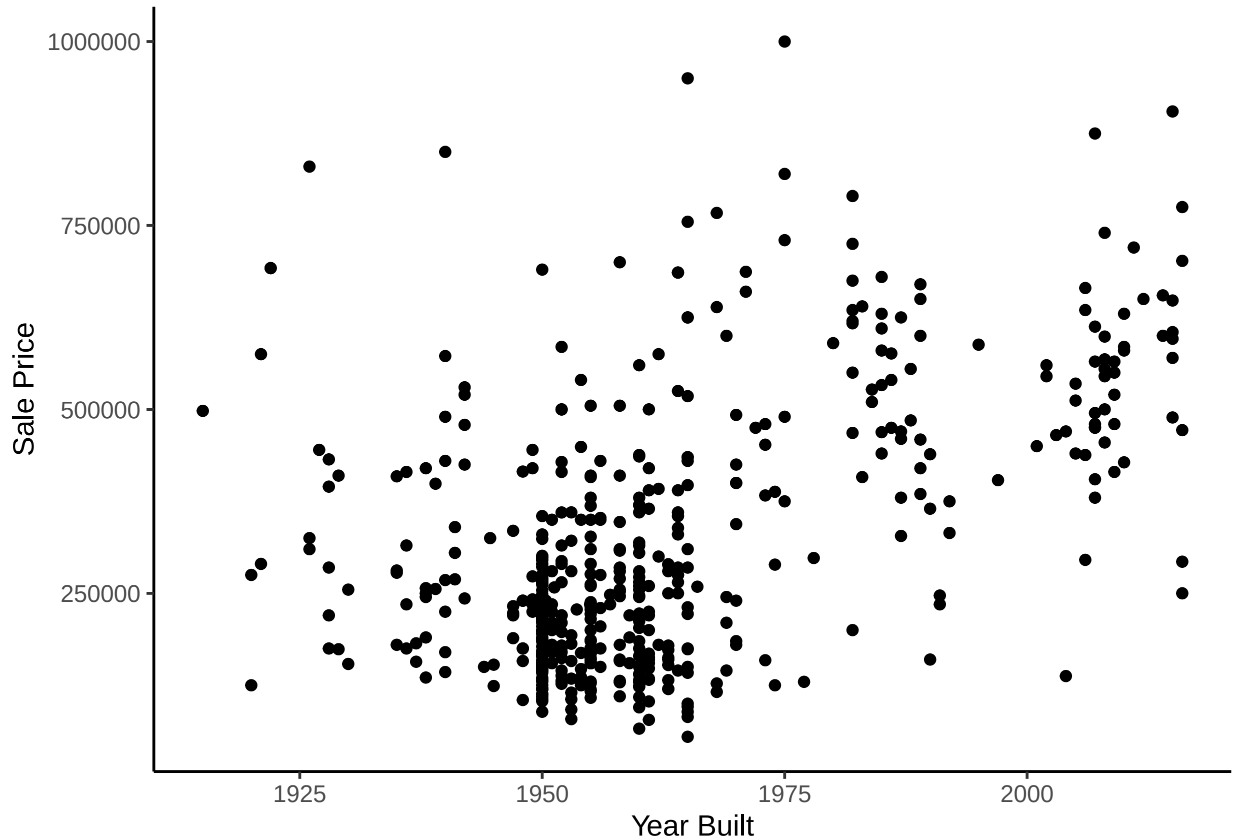

```
ggplot(df_final, aes(x=walk_score, y=sale_price)) + geom_point() +  theme_classic()  +
  labs(
       x="Walk Score", y = "Sale Price")
```



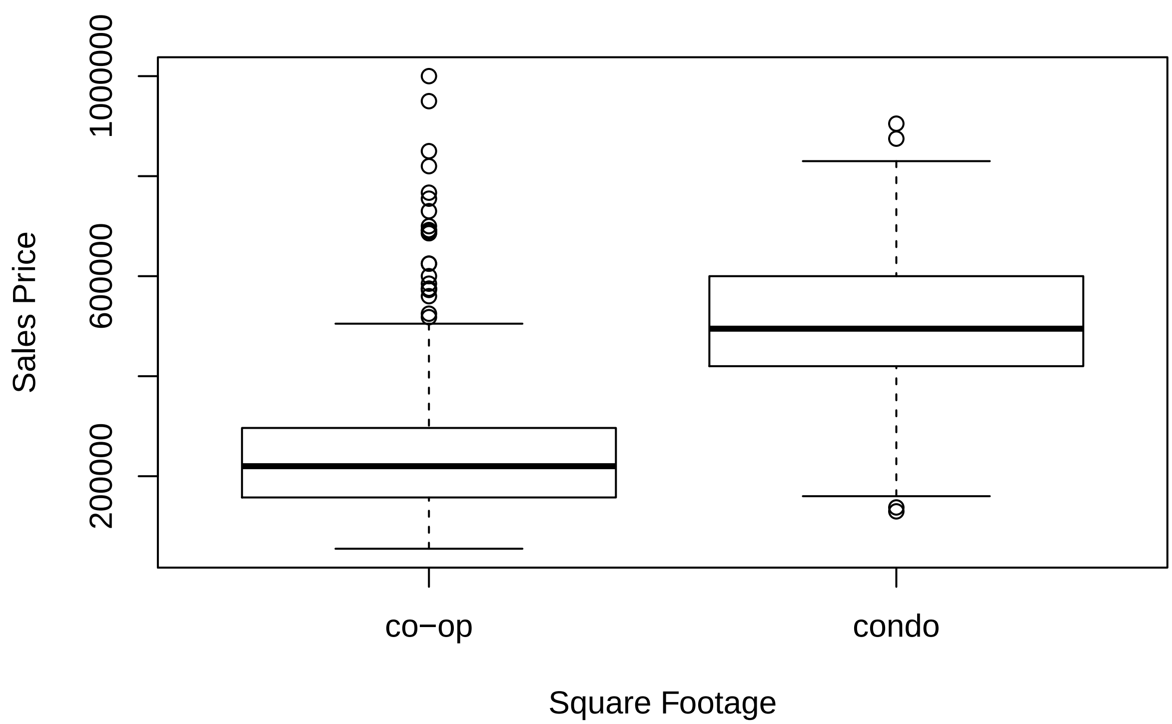

```
ggplot(df_final, aes(x=common_charges, y=sale_price)) + geom_point() +  theme_classic()  +
  labs(
       x="Common Charges", y = "Sale Price")
```




```
ggplot(df_final, aes(x=approx_year_built, y=sale_price)) + geom_point() +  theme_classic()   +
  labs(
        x="Year Built", y = "Sale Price")
```




```
plot(y=df_clean$sale_price,df_clean$coop_condo, xlab ='Square Footage', ylab= 'Sales Price',)
```



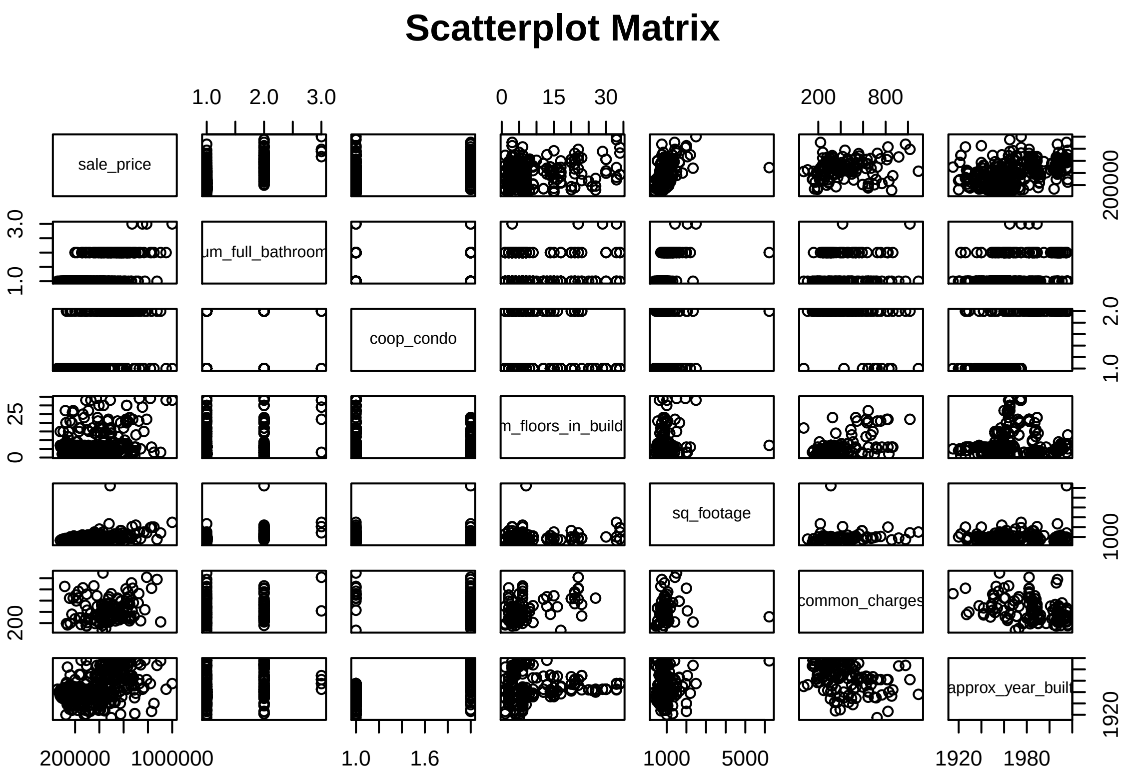

```
# Tried to one hot incode data. MAJOR FAIL. R takes care of this since data is already factors
# df_dummy = copy(df_final)
```

```r
# df_dummy$cats_allowed = model.matrix(~df_dummy$cats_allowed + 0)
# df_dummy$coop_condo = model.matrix(~df_dummy$coop_condo + 0)
# df_dummy$dining_room_type = model.matrix(~df_dummy$dining_room_type + 0)
# df_dummy$dogs_allowed = model.matrix(~df_dummy$dogs_allowed + 0)
# df_dummy$fuel_type = model.matrix(~df_dummy$fuel_type + 0)
# library(data.table,mltools)
# something = copy(df_final)
#
# something$fuel_type = cbind(model.matrix(~something$fuel_type))
```

```r
colnames(df_final)
```

```
##  [1] "cats_allowed"                       "common_charges"
##  [3] "coop_condo"                         "dining_room_type"
##  [5] "dogs_allowed"                       "fuel_type"
##  [7] "garage_exists"                      "maintenance_cost"
##  [9] "approx_year_built"                  "num_bedrooms"
## [11] "num_floors_in_building"             "num_full_bathrooms"
## [13] "num_total_rooms"                    "sq_footage"
## [15] "sale_price"                         "walk_score"
## [17] "is_missing_common_charges"          "is_missing_dining_room_type"
## [19] "is_missing_maintenance_cost"        "is_missing_approx_year_built"
## [21] "is_missing_num_floors_in_building" "is_missing_sq_footage"
```

```r
pairs(~sale_price+num_full_bathrooms+coop_condo+num_floors_in_building+sq_footage+common_charges+approx_
   main="Scatterplot Matrix")
```




```r
skim(df_final)
```

Table 13: Data summary

Name	df_final
Number of rows	528
Number of columns	22
Column type frequency:	
factor	5
numeric	17
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
cats_allowed	0	1	FALSE	2	no: 285, yes: 243
coop_condo	0	1	FALSE	2	co-: 399, con: 129
dining_room_type	0	1	FALSE	4	com: 332, for: 139, oth: 55, din: 2
dogs_allowed	0	1	FALSE	2	no: 381, yes: 147
fuel_type	0	1	FALSE	4	gas: 301, oil: 180, oth: 36, ele: 11

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p
common_charges	0	1	489.07	138.73	70	417.43	469.
garage_exists	0	1	0.18	0.38	0	0.00	0.
maintenance_cost	0	1	810.71	361.55	155	605.50	722.
approx_year_built	0	1	1962.25	20.48	1915	1950.00	1956.
num_bedrooms	0	1	1.54	0.75	0	1.00	1.
num_floors_in_building	0	1	7.08	6.33	1	3.00	6.
num_full_bathrooms	0	1	1.20	0.42	1	1.00	1.
num_total_rooms	0	1	4.02	1.20	1	3.00	4.
sq_footage	0	1	901.76	364.21	375	722.87	835.
sale_price	0	1	314956.56	179526.60	55000	171500.00	259500.
walk_score	0	1	83.10	13.09	15	76.00	85.
is_missing_common_charges	0	1	0.75	0.43	0	0.75	1.
is_missing_dining_room_type	0	1	0.23	0.42	0	0.00	0.
is_missing_maintenance_cost	0	1	0.27	0.44	0	0.00	0.
is_missing_approx_year_built	0	1	0.01	0.11	0	0.00	0.
is_missing_num_floors_in_building	0	1	0.20	0.40	0	0.00	0.
is_missing_sq_footage	0	1	0.60	0.49	0	0.00	1.

```r
set.seed(28)
train.control <- trainControl(method = "repeatedcv",
                              number = 10, repeats = 3)
prop_test = 0.10
test_indices = sample(1 : nrow(df_final), round((prop_test) * nrow(df_final)))
df_test = df_final[test_indices, ]
y_test = df_test$sale_price
X_test = cbind(1, df_test)
```

```r
#X_test$sale_price = NULL


train_indices = setdiff(1 : nrow(df_final), test_indices)
df_train = df_final[train_indices, ]
y_train = df_train$sale_price
X_train = cbind(1, df_train)
#X_train$sale_price = NULL

n_train = nrow(X_train)
```

```r
#mod = train(sale_price ~ ., df_final, trControl = train.control,method = "lm")
mod = lm(sale_price ~ ., df_final)
summary(mod)$r.squared
```

```
## [1] 0.77106
```

```r
summary(mod)$sigma
```

```
## [1] 88012.24
```

```r
summary(mod)
```

```
##
## Call:
## lm(formula = sale_price ~ ., data = df_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -349600  -49834   -2811   41151  354828
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                      -155672.69  630147.20  -0.247
## cats_allowedyes                     8259.94   10904.47   0.757
## common_charges                        41.95      46.23   0.907
## coop_condocondo                   269327.88   24446.08  11.017
## dining_room_typedining area        18414.34   63357.85   0.291
## dining_room_typeformal             13871.81   10362.41   1.339
## dining_room_typeother              23839.44   13955.36   1.708
## dogs_allowedyes                    16391.72   12111.28   1.353
## fuel_typegas                        5820.46   28113.12   0.207
## fuel_typeoil                        4870.13   28854.99   0.169
## fuel_typeother                     25405.43   31956.20   0.795
## garage_exists                       2612.47   11269.62   0.232
## maintenance_cost                     106.06      20.81   5.096
## approx_year_built                    -45.63     321.37  -0.142
## num_bedrooms                       57695.80    9432.52   6.117
## num_floors_in_building              5329.85     856.38   6.224
## num_full_bathrooms                 74247.49   13549.78   5.480
## num_total_rooms                       63.03    6248.70   0.010
## sq_footage                            27.21      15.75   1.728
## walk_score                          1299.14     316.12   4.110
## is_missing_common_charges          38818.06   24560.10   1.581
## is_missing_dining_room_type         8535.89    9581.08   0.891
```

```
## is_missing_maintenance_cost        -28263.28   22039.47  -1.282
## is_missing_approx_year_built        20074.99   37107.74   0.541
## is_missing_num_floors_in_building   19195.20   10071.58   1.906
## is_missing_sq_footage              -17027.67    8413.31  -2.024
##                                               Pr(>|t|)
## (Intercept)                                     0.8050
## cats_allowedyes                                 0.4491
## common_charges                                  0.3646
## coop_condocondo                   < 0.0000000000000002 ***
## dining_room_typedining area                     0.7714
## dining_room_typeformal                          0.1813
## dining_room_typeother                           0.0882 .
## dogs_allowedyes                                 0.1765
## fuel_typegas                                    0.8361
## fuel_typeoil                                    0.8660
## fuel_typeother                                  0.4270
## garage_exists                                   0.8168
## maintenance_cost                        0.00000049094 ***
## approx_year_built                               0.8871
## num_bedrooms                            0.00000000192 ***
## num_floors_in_building                  0.00000000103 ***
## num_full_bathrooms                      0.00000006756 ***
## num_total_rooms                                 0.9920
## sq_footage                                      0.0846 .
## walk_score                              0.00004627702 ***
## is_missing_common_charges                       0.1146
## is_missing_dining_room_type                     0.3734
## is_missing_maintenance_cost                     0.2003
## is_missing_approx_year_built                    0.5888
## is_missing_num_floors_in_building               0.0572 .
## is_missing_sq_footage                           0.0435 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88010 on 502 degrees of freedom
## Multiple R-squared:  0.7711, Adjusted R-squared:  0.7597
## F-statistic: 67.63 on 25 and 502 DF,  p-value: < 0.00000000000000022
```

```r
set.seed(28)
mod =lm(sale_price ~., data.frame(df_train),set.seed(28))
summary(mod)$r.squared
```

```
## [1] 0.7649951
```

```r
summary(mod)$sigma
```

```
## [1] 88712.26
```

```r
y_hat = predict(mod,data.frame(X_test))
e = y_test - y_hat
Rsq_oos = (var(y_test) - var(e)) / var(y_test)

cat("My R Squared in sample is ",summary(mod)$r.squared, "My RSME is:", summary(mod)$sigma)
```

```
## My R Squared in sample is  0.7649951 My RSME is: 88712.26
```

```r
cat("\nMy R Squared out of sample is ",Rsq_oos, "My RSME is:", sd(e))
```

```
##
## My R Squared out of sample is  0.8046578 My RSME is: 85564.51
```

```r
#plot(y_test,y_hat)
summary(mod)
```

```
##
## Call:
## lm(formula = sale_price ~ ., data = data.frame(df_train), subset = set.seed(28))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -353328  -46131   -1660   43358  353257
##
## Coefficients:
##                                Estimate Std. Error t value
## (Intercept)                  -298395.10  659717.40  -0.452
## cats_allowedyes                 9080.70   11575.97   0.784
## common_charges                    31.45      52.67   0.597
## coop_condocondo               268580.86   25419.10  10.566
## dining_room_typedining area     23700.90   89632.03   0.264
## dining_room_typeformal          7681.67   11135.26   0.690
## dining_room_typeother          32171.80   14691.74   2.190
## dogs_allowedyes                19982.52   12690.44   1.575
## fuel_typegas                   -3832.21   29743.63  -0.129
## fuel_typeoil                   -4827.69   30620.42  -0.158
## fuel_typeother                  5305.87   33823.15   0.157
## garage_exists                    802.17   12224.26   0.066
## maintenance_cost                 121.29      26.87   4.514
## approx_year_built                 33.66     336.42   0.100
## num_bedrooms                   54268.57   10137.47   5.353
## num_floors_in_building          4939.70     918.37   5.379
## num_full_bathrooms             82887.84   14540.51   5.700
## num_total_rooms                  378.61    6557.42   0.058
## sq_footage                        23.79      16.32   1.458
## walk_score                      1134.48     336.43   3.372
## is_missing_common_charges      42305.48   25406.07   1.665
## is_missing_dining_room_type     7221.66   10162.44   0.711
## is_missing_maintenance_cost   -26899.32   22966.64  -1.171
## is_missing_approx_year_built   23817.49   37540.65   0.634
## is_missing_num_floors_in_building  14964.93   10575.50   1.415
## is_missing_sq_footage         -15340.57    8925.11  -1.719
##                                           Pr(>|t|)
## (Intercept)                               0.651266
## cats_allowedyes                           0.433193
## common_charges                            0.550712
## coop_condocondo              < 0.0000000000000002 ***
## dining_room_typedining area               0.791574
## dining_room_typeformal                    0.490644
## dining_room_typeother                     0.029053 *
## dogs_allowedyes                           0.116050
## fuel_typegas                              0.897541
```

```
## fuel_typeoil                                    0.874794
## fuel_typeother                                  0.875417
## garage_exists                                   0.947709
## maintenance_cost                            0.0000081193 ***
## approx_year_built                               0.920351
## num_bedrooms                                0.0000001381 ***
## num_floors_in_building                      0.0000001209 ***
## num_full_bathrooms                          0.0000000217 ***
## num_total_rooms                                 0.953983
## sq_footage                                      0.145619
## walk_score                                      0.000811 ***
## is_missing_common_charges                       0.096576 .
## is_missing_dining_room_type                     0.477687
## is_missing_maintenance_cost                     0.242126
## is_missing_approx_year_built                    0.526113
## is_missing_num_floors_in_building               0.157745
## is_missing_sq_footage                           0.086338 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88710 on 449 degrees of freedom
## Multiple R-squared:  0.765,  Adjusted R-squared:  0.7519
## F-statistic: 58.46 on 25 and 449 DF,  p-value: < 0.00000000000000022
```

```
# pacman::p_load(ggplot2, mlr3,mlr)
# library(mlr3)
# library(mlr)
# modeling_task = makeRegrTask(data=data.frame(X_train), target='sale_price')
# algorithm = makeLearner("regr.lm")
# validation = makeResampleDesc("CV", iters = 5)
# #Having issues with fuel_type none
# res = resample(algorithm, modeling_task, validation,measures = list(rmse))
# res
# #average rsme somehow worse than above
# mean(res$measures.test$rmse)
```

REGRESSION TREEES.

Here the trees overfit in sample but they did pretty decent out of sample but not better than OLS

```
options(java.parameters = "-Xmx4000m")

X_train_CART = X_train
X_train_CART$sale_price = NULL

X_test_CART = X_test
X_test_CART$sale_price = NULL
tree_model = YARFCART(X_train_CART, y_train, bootstrap_indices = 1 : n_train, calculate_oob_error = TRUE
```

```
## YARF initializing with a fixed 1 trees...
## YARF factors created...
## YARF after data preprocessed... 31 total features...
## Beginning YARF regression model construction...done.
## Calculating OOB error...done.
```

```r
illustrate_trees(tree_model, max_depth = 4, open_file = TRUE,margin_in_px=200,font_size=20,length_in_px
get_tree_num_nodes_leaves_max_depths(tree_model)
```

```
## $num_nodes
## [1] 375
##
## $num_leaves
## [1] 188
##
## $max_depths
## [1] 25
```

```r
#In Sample Error
y_hat_train = predict(tree_model,X_train)
```

```
## Warning in predict.YARF(tree_model, X_train): Prediction set column names did not match training set
## Attempting to subset to training set columns.
```

```r
e_train = y_train - y_hat_train
rsme_train = sd(e_train)
rsquared_train = (var(y_train) - var(e_train)) / var(y_train)

#Out of Sample Error
y_hat_test = predict(tree_model, X_test)
```

```
## Warning in predict.YARF(tree_model, X_test): Prediction set column names did not match training set c
## Attempting to subset to training set columns.
```

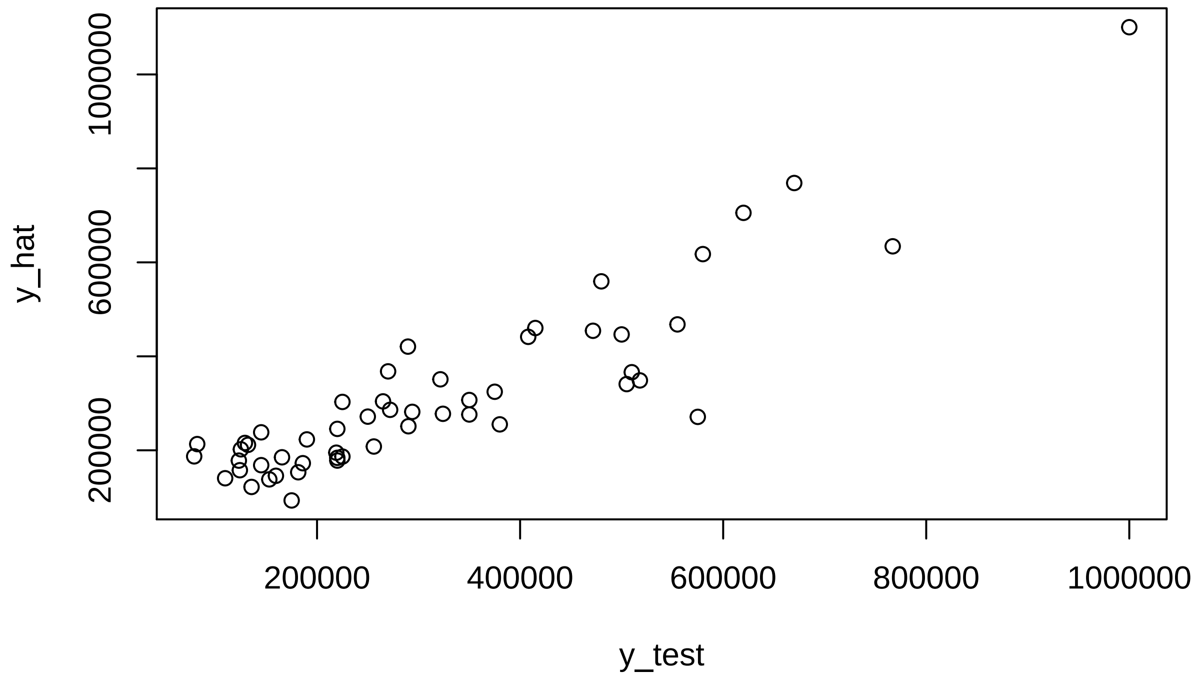
```r
e_test = y_test - y_hat_test
rsme_test = sd(e_test)
rsquared_test = (var(y_test) - var(e_test)) / var(y_test)

cat("My R Squared in sample is ",rsquared_train, "My RSME is:", rsme_train)
```

```
## My R Squared in sample is  0.9882321 My RSME is: 19320.98
```

```r
cat("\nMy R Squared out of sample is ",rsquared_test, "My RSME is:",rsme_test)
```

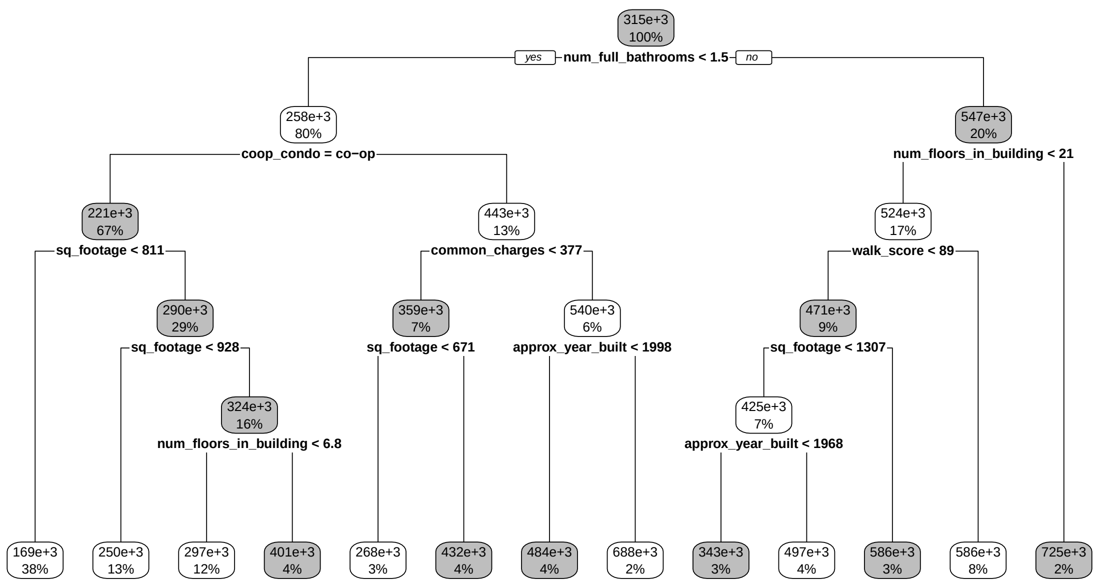
```
##
## My R Squared out of sample is  0.7961257 My RSME is: 87413.18
```

```r
plot(y_test,y_hat)
```




```r
library(rpart,mlr)
library("rpart.plot")
rpart_fit = rpart(sale_price ~., data.frame(X_train),method="anova")
rpart.plot(rpart_fit,box.col=c("grey", "white"))
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a matrix).
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```




```r
rpart_fit
```

```
## n= 475
##
## node), split, n, deviance, yval
```

```
##       * denotes terminal node
##
##  1) root 475 15036150000000 314877.8
##    2) num_full_bathrooms< 1.5 381  6589647000000 257531.4
##      4) coop_condo=co-op 318  2568349000000 220853.0
##        8) sq_footage< 811.1617 181   389403400000 168632.4 *
##        9) sq_footage>=811.1617 137  1033253000000 289845.1
##         18) sq_footage< 928.3983 63   336600400000 249738.1 *
##         19) sq_footage>=928.3983 74   509036800000 323990.2
##           38) num_floors_in_building< 6.765 55   307402000000 297298.0 *
##           39) num_floors_in_building>=6.765 19    49014940000 401257.3 *
##      5) coop_condo=condo 63  1434082000000 442670.1
##       10) common_charges< 377 34   494643300000 359439.4
##         20) sq_footage< 671.3365 15    92026640000 267975.9 *
##         21) sq_footage>=671.3365 19   178067100000 431647.4 *
##       11) common_charges>=377 29   427770100000 540250.9
##         22) approx_year_built< 1997.5 21   111454700000 483780.4 *
##         23) approx_year_built>=1997.5 8    73558990000 688486.0 *
##    3) num_full_bathrooms>=1.5 94  2115051000000 547313.8
##      6) num_floors_in_building< 20.5 83  1618691000000 523704.8
##       12) walk_score< 88.5 45  1045441000000 471400.0
##         24) sq_footage< 1307.457 32   432133500000 425031.2
##           48) approx_year_built< 1967.5 15    98365230000 343366.7 *
##           49) approx_year_built>=1967.5 17   145464100000 497088.2 *
##         25) sq_footage>=1307.457 13   375147200000 585538.5 *
##       13) walk_score>=88.5 38   304349500000 585644.7 *
##      7) num_floors_in_building>=20.5 11   101022700000 725454.5 *
```

```r
library(mlr,mlr3)
```

```
## Loading required package: ParamHelpers

## 'mlr' is in maintenance mode since July 2019. Future development
## efforts will go into its successor 'mlr3' (<https://mlr3.mlr-org.com>).

##
## Attaching package: 'mlr'

## The following object is masked from 'package:caret':
##
##     train
```

```r
modeling_task = makeRegrTask(data=data.frame(X_train), target='sale_price')
```

```
## Warning in makeTask(type = type, data = data, weights = weights, blocking =
## blocking, : Empty factor levels were dropped for columns: dining_room_type
```

```r
algorithm = makeLearner("regr.rpart")
validation = makeResampleDesc("CV", iters = 5)
#Having issues with fuel_type none
res = resample(algorithm, modeling_task, validation,measures = list(rmse))
```

```
## Resampling: cross-validation

## Measures:             rmse

## [Resample] iter 1:    77933.2398912

## [Resample] iter 2:    98589.0648772
```

```
## [Resample] iter 3:    99565.4579709

## [Resample] iter 4:    113289.5639943

## [Resample] iter 5:    106979.1754515

##

## Aggregated Result: rmse.test.rmse=99985.7422881

##
```

```
res
```

```
## Resample Result
## Task: data.frame(X_train)
## Learner: regr.rpart
## Aggr perf: rmse.test.rmse=99985.7422881
## Runtime: 0.069572
```

```
#average rsme somehow worse than above
mean(res$measures.test$rmse)
```

```
## [1] 99271.3
```

```
X_train_RF = X_train
X_train_RF$sale_price = NULL

X_test_RF = X_test
X_test_RF$sale_price = NULL
Bag_model = YARFBAG(X_train_RF, y_train, num_trees = 250, seed = 1 ,calculate_oob_error = TRUE)
```

```
## YARF initializing with a fixed 250 trees...
## YARF factors created...
## YARF after data preprocessed... 31 total features...
## Beginning YARF regression model construction...done.
## Calculating OOB error...done.
```

```
Bag_model$rmse_oob
```

```
## [1] 77425.65
```

```
y_hat_test_bag = predict(Bag_model, X_test)
```

```
## Warning in predict.YARF(Bag_model, X_test): Prediction set column names did not match training set c
## Attempting to subset to training set columns.
```

```
s_e_bag = sd(y_test - y_hat_test_bag)
s_e_bag
```

```
## [1] 76296.75
```

```
#In Sample Error
y_hat_train = predict(Bag_model,X_train)
```

```
## Warning in predict.YARF(Bag_model, X_train): Prediction set column names did not match training set
## Attempting to subset to training set columns.
```

```
e_train = y_train - y_hat_train
rsme_train = sd(e_train)
rsquared_train = (var(y_train) - var(e_train)) / var(y_train)
```

```
#Out of Sample Error
y_hat_test = predict(Bag_model, X_test)
```

```
## Warning in predict.YARF(Bag_model, X_test): Prediction set column names did not match training set co
## Attempting to subset to training set columns.
```

```
e_test = y_test - y_hat_test
rsme_test = sd(e_test)
rsquared_test = (var(y_test) - var(e_test)) / var(y_test)

cat("My R Squared in sample is ",rsquared_train, "My RSME is:", rsme_train,"\n" )
```

```
## My R Squared in sample is  0.9702822 My RSME is: 30703.46
```

```
cat("\nMy R Squared out of sample is ",rsquared_test, "My RSME is:",rsme_test,"\n" )
```

```
##
## My R Squared out of sample is  0.8446824 My RSME is: 76296.75
```

```
df_final_bag_all = copy(df_final)
y_all = df_final_bag_all$sale_price
df_final_bag_all$sale_price = NULL
mod_bag_all = YARFBAG(df_final_bag_all, y_all, num_trees = 250, seed = 28)
```

```
## YARF initializing with a fixed 250 trees...
## YARF factors created...
## YARF after data preprocessed... 30 total features...
## Beginning YARF regression model construction...done.
## Calculating OOB error...done.
```

```
mod_bag_all$rmse_oob
```

```
## [1] 76421.52
```

RANDOM FOREST

```
set.seed(2)
X_train_RF = X_train
X_train_RF$sale_price = NULL

X_test_RF = X_test
X_test_RF$sale_price = NULL

RF_model = YARF(X_train_RF, y_train, num_trees = 250, seed = 1 ,calculate_oob_error = TRUE)
```

```
## YARF initializing with a fixed 250 trees...
## YARF factors created...
## YARF after data preprocessed... 31 total features...
## Beginning YARF regression model construction...done.
## Calculating OOB error...done.
```

```
#In Sample Error
y_hat_train = predict(RF_model,X_train)
```

```
## Warning in predict.YARF(RF_model, X_train): Prediction set column names did not match training set co
## Attempting to subset to training set columns.
```

```r
e_train = y_train - y_hat_train
rsme_train = sd(e_train)
rsquared_train = (var(y_train) - var(e_train)) / var(y_train)

#Out of Sample Error
y_hat_test = predict(RF_model, X_test)
```

```
## Warning in predict.YARF(RF_model, X_test): Prediction set column names did not match training set col
## Attempting to subset to training set columns.
```

```r
e_test = y_test - y_hat_test
rsme_test = sd(e_test)
rsquared_test = (var(y_test) - var(e_test)) / var(y_test)

cat("My R Squared in sample is ",rsquared_train, "My RSME is:", rsme_train,"\n" )
```

```
## My R Squared in sample is  0.9657972 My RSME is: 32938.96
```

```r
cat("\nMy R Squared out of sample is ",rsquared_test, "My RSME is:",rsme_test,"\n" )
```

```
##
## My R Squared out of sample is  0.820339 My RSME is: 82058.32
```

```r
cat("OOB RSME:",RF_model$rmse_oob,"\n" )
```

```
## OOB RSME: 78439.13
```

```r
cat("GAIN OVER TREES",(mod_bag_all$rmse_oob - RF_model$rmse_oob) / mod_bag_all$rmse_oob * 100, "%\n" )
```

```
## GAIN OVER TREES -2.640113 %
```

```r
RF_model$rmse_oob
```

```
## [1] 78439.13
```

```r
set.seed(28)
modeling_task = makeRegrTask(data=data.frame(X_train), target='sale_price')
```

```
## Warning in makeTask(type = type, data = data, weights = weights, blocking =
## blocking, : Empty factor levels were dropped for columns: dining_room_type
```

```r
algorithm = makeLearner("regr.randomForest")
validation = makeResampleDesc("CV", iters = 5)
#Having issues with fuel_type none
res = resample(algorithm, modeling_task, validation,measures = list(rmse))
```

```
## Resampling: cross-validation

## Measures:             rmse

## [Resample] iter 1:    87545.1912787

## [Resample] iter 2:    75434.2169342

## [Resample] iter 3:    73396.4503821

## [Resample] iter 4:    82743.6328139

## [Resample] iter 5:    55317.9247182

##

## Aggregated Result: rmse.test.rmse=75694.2561831
```

```
##
```

```
res
```

```
## Resample Result
## Task: data.frame(X_train)
## Learner: regr.randomForest
## Aggr perf: rmse.test.rmse=75694.2561831
## Runtime: 2.80878
```

```
#average rsme somehow worse than above
mean(res$measures.test$rmse)
```

```
## [1] 74887.48
```

```
set.seed(28)
modeling_task = makeRegrTask(data=data.frame(X_train), target='sale_price')
```

```
## Warning in makeTask(type = type, data = data, weights = weights, blocking =
## blocking, : Empty factor levels were dropped for columns: dining_room_type
```

```
algorithm = makeLearner("regr.randomForest")
holdout = makeResampleDesc("Holdout")
#Having issues with fuel_type none
res = resample(algorithm, modeling_task, holdout,measures = list(rmse))
```

```
## Resampling: holdout
```

```
## Measures:             rmse
```

```
## [Resample] iter 1:    85520.2547615
```

```
##
```

```
## Aggregated Result: rmse.test.rmse=85520.2547615
```

```
##
```

```
res
```

```
## Resample Result
## Task: data.frame(X_train)
## Learner: regr.randomForest
## Aggr perf: rmse.test.rmse=85520.2547615
## Runtime: 0.40167
```

```
#average rsme somehow worse than above
mean(res$measures.test$rmse)
```

```
## [1] 85520.25
```

```
# library(rpart)
# library(rpart.plot)
# fit = rpart(sale_price ~., data.frame(X_train),method="anova")
# rpart.plot(fit)
# summary(fit)
# pred
#
# in_e = y_train - pred
# sd(in_e)
# (var(y_train) - var(e)) / var(y_train)
# e = y_test - pred
```

```
# sd(e)
#
# Rsq_oos = (var(y_test) - var(e)) / var(y_test)
# #sd(e)
# cat("My R Squared in sample is ",summary(mod)$r.squared, "My RSME is:", sd(in_e))
# cat("\nMy R Squared out of sample is ",Rsq_oos, "My RSME is:", sd(e))
# ```
# ```{r}
# library(randomForest)
# control <- trainControl(method="cv", number=10)
#
#
# RegressionTree1 = train(sale_price~., data=data.frame(X_train), method="rpart", trControl=control)
# y_hat = predict(object = RegressionTree1,newdata = data.frame(X_test))
# sqrt(mean((y_hat-y_test)^2))
#
# RegressionTree = train(sale_price~., data=df_final, method="rpart", trControl=control)
# print(RegressionTree)
#
#
# ##
#
# fit = rpart(sale_price ~., data.frame(X_train),method = 'anova')
# printcp(fit)
# rpart.plot(fit)
# summary(fit)
# y_hat = predict(object = fit,newdata = data.frame(X_test))
# sqrt(mean((y_hat-y_test)^2))
```

```
# RandomForest = train(sale_price~., data=df_final, method="rf", trControl=control)
# print(RandomForest)
#
#
```