

##	parcelid	logerror	date
##	Min. : 10711738	Min. :-4.60500	2016-07-29: 910
##	1st Qu.: 11559500	1st Qu.: -0.02530	2016-04-29: 902
##	Median : 12547337	Median : 0.00600	2016-09-30: 894
##	Mean : 12984656	Mean : 0.01146	2016-06-30: 874
##	3rd Qu.: 14227552	3rd Qu.: 0.03920	2016-05-27: 863

```
## Max.      :162960842    Max.      : 4.73700    2016-08-31: 737
##                                     (Other)    :85095
```

head of transactions data set

```
##   parcelid logerror      date
## 1 11016594   0.0276 2016-01-01
## 2 14366692  -0.1684 2016-01-01
## 3 12098116  -0.0040 2016-01-01
## 4 12643413   0.0218 2016-01-02
## 5 14432541  -0.0050 2016-01-02
## 6 11509835  -0.2705 2016-01-02
```

head of properties data set

```
##           X parcelid aircon architectural_style area_basement num_bathroom
## 1 2342182 12484350      1                      NA             NA           4
## 2  541880 12181530      NA                      NA             NA           1
## 3 1617969 14222705      NA                      NA             NA           0
## 4 1729095 11467838      NA                      NA             NA           2
## 5   84189 13015665      NA                      NA             NA           1
## 6 1570086 12633721      NA                      NA             NA           4
##   num_bedroom framing quality num_bathroom_calc deck area_firstfloor_finis
##   hed
## 1           6      NA      4                4  NA
## NA
## 2           2      NA      7                1  NA
## NA
## 3           0      NA      NA               NA  NA
## NA
## 4           3      NA      7                2  NA
## NA
## 5           2      NA      7                1  NA
## NA
## 6           6      NA      7                4  NA
## NA
##   area_total_calc area_live_finished area_liveperi_finished area_total_fin
##   ished
## 1           3315                3315                NA
## NA
## 2           1088                1088                NA
## NA
## 3           4231                 NA                NA
## NA
## 4           1327                1327                NA
## NA
## 5           976                 976                NA
## NA
## 6           3030                 NA                NA
## 3030
##   area_unknown area_base fips num_fireplace num_bath num_garage area_garag
```

e							
## 1	NA	NA 6037	NA	4	NA	N	
A							
## 2	NA	NA 6037	NA	1	NA	N	
A							
## 3	NA	4231 6059	NA	NA	0		
0							
## 4	NA	NA 6037	NA	2	NA	N	
A							
## 5	NA	NA 6037	NA	1	NA	N	
A							
## 6	NA	NA 6037	NA	4	NA	N	
A							
##	flag_tub	heating	latitude	longitude	area_lot	num_pool	area_pool pooltyp
eid10							
## 1		2	33852329	-118125098	6548	NA	NA
NA							
## 2		NA	33986932	-118297793	6008	NA	NA
NA							
## 3		NA	33819470	-117832954	9118	NA	NA
NA							
## 4		7	33958756	-118398254	5145	NA	NA
NA							
## 5		7	34138321	-117908275	7000	NA	NA
NA							
## 6		NA	33782587	-118251435	7000	NA	NA
NA							
##	pooltypeid2	pooltypeid7	zoning_landuse	county	zoning_landuse	zoning_prop	erty
## 1	NA	NA		0100		261	LK
R1YY							
## 2	NA	NA		0100		261	
LAR2							
## 3	NA	NA		96		248	
## 4	NA	NA		0100		261	
LAR1							
## 5	NA	NA		0100		261	AZ
R1C*							
## 6	NA	NA		0200		246	
LAR2							
##	rawcensustractandblock	region_city	region_county	region_neighbor	region_	zip	
## 1		60375708	12292	3101		NA	96
212							
## 2		60372372	12447	3101		118208	96
025							
## 3		60590758	33252	1286		NA	97
063							
## 4		60372780	12447	3101		7877	96
026							

```

## 5          60374006          37015          3101          NA          96
464
## 6          60372947          12447          3101          48516          96
228
##  num_room story num_75_bath material num_unit area_patio area_shed build_
year
## 1          0     NA          NA          NA          1          NA          NA
1950
## 2          0     NA          NA          NA          1          NA          NA
1940
## 3          0     NA          NA          NA          4          NA          NA
1973
## 4          0     NA          NA          NA          1          NA          NA
1944
## 5          0     NA          NA          NA          1          NA          NA
1921
## 6          0     NA          NA          NA          2          NA          NA
1921
##  num_story flag_fireplace tax_building tax_total tax_year tax_land
## 1          NA          255065    277073    2015    22008
## 2          NA          22500    172568    2015    150068
## 3          2          212125    353550    2015    141425
## 4          NA          97880    370695    2015    272815
## 5          NA          94231    262698    2015    168467
## 6          NA          225985    242687    2015    16702
##  tax_property tax_delinquency tax_delinquency_year censustractandblock
## 1    3766.44          NA          6.037571e+13
## 2    2186.02          NA          6.037237e+13
## 3    4655.98          NA          6.059076e+13
## 4    4610.51          NA          6.037278e+13
## 5    3303.92          NA          6.037401e+13
## 6    3144.99          NA          6.037295e+13

```

Column names

properties

X, parcelid, aircon, architectural_style, area_basement, num_bathroom, num_bedroom, framing, quality, num_bathroom_calc, deck, area_firstfloor_finished, area_total_calc, area_live_finished, area_liveperi_finished, area_total_finished, area_unknown, area_base, fips, num_fireplace, num_bath, num_garage, area_garage, flag_tub, heating, latitude, longitude, area_lot, num_pool, area_pool, pooltypeid10, pooltypeid2, pooltypeid7, zoning_landuse_county, zoning_landuse, zoning_property, rawcensustractandblock, region_city, region_county, region_neighbor, region_zip, num_room, story, num_75_bath, material, num_unit, area_patio, area_shed, build_year, num_story, flag_fireplace, tax_building, tax_total, tax_year, tax_land, tax_property, tax_delinquency, tax_delinquency_year, censustractandblock

transactions

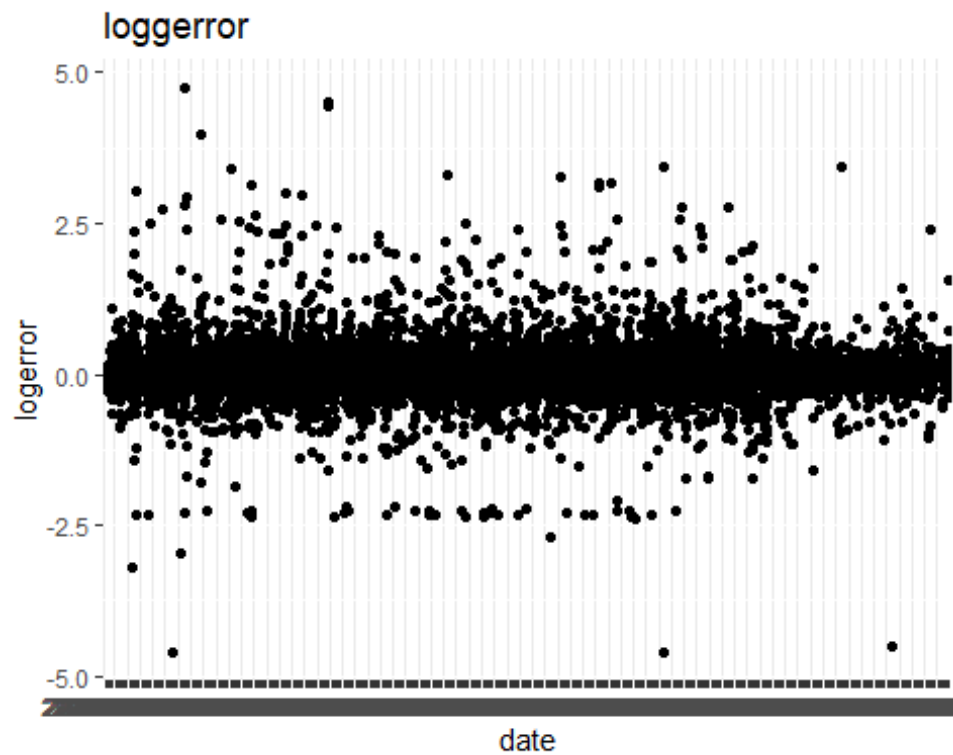
parcelid, logerror, date

statistics of logerror in transactions

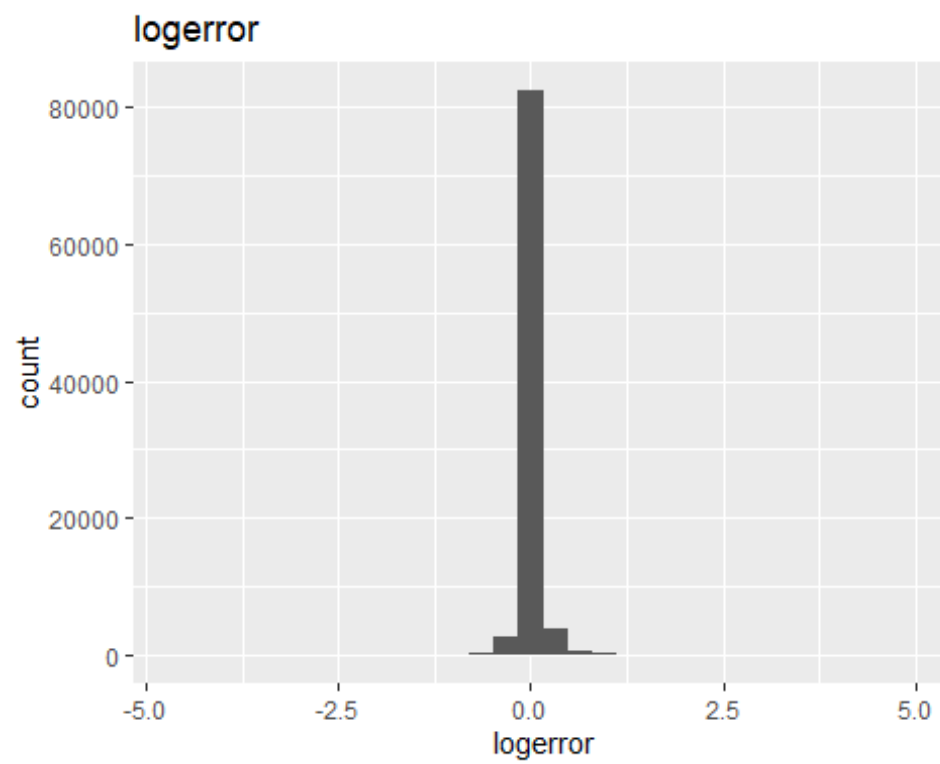
- Mean : 0.0114572
- median : 0.006
- std : ' 0.1610788
- Max: 4.737
- Min: -4.605

Data exploration

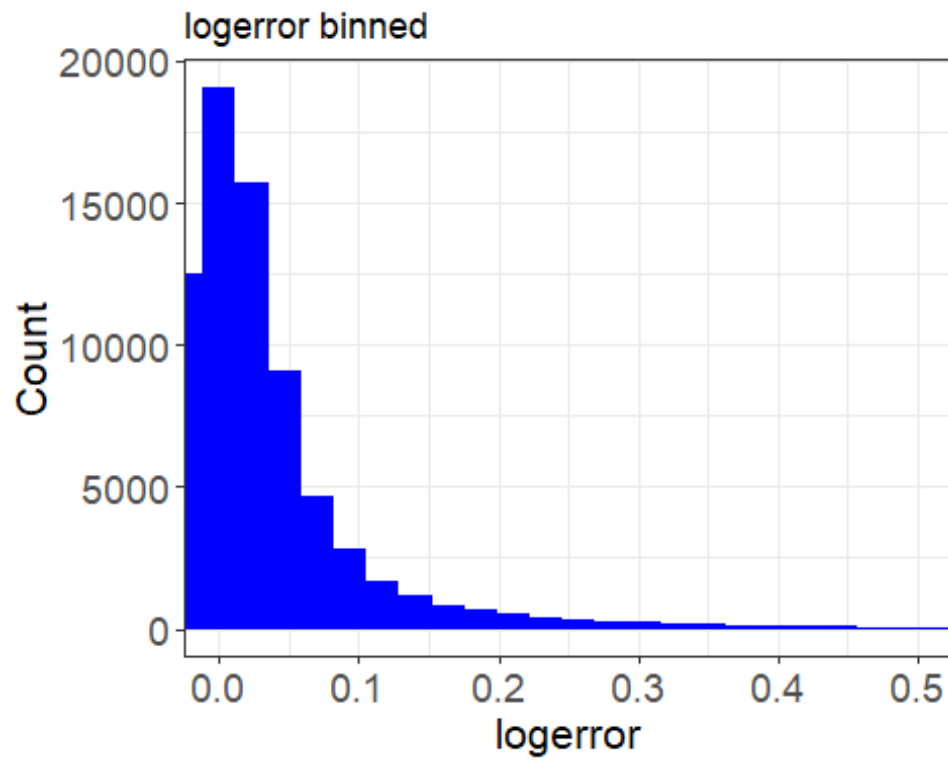
Scatter plot of logerror



histogram of logerror

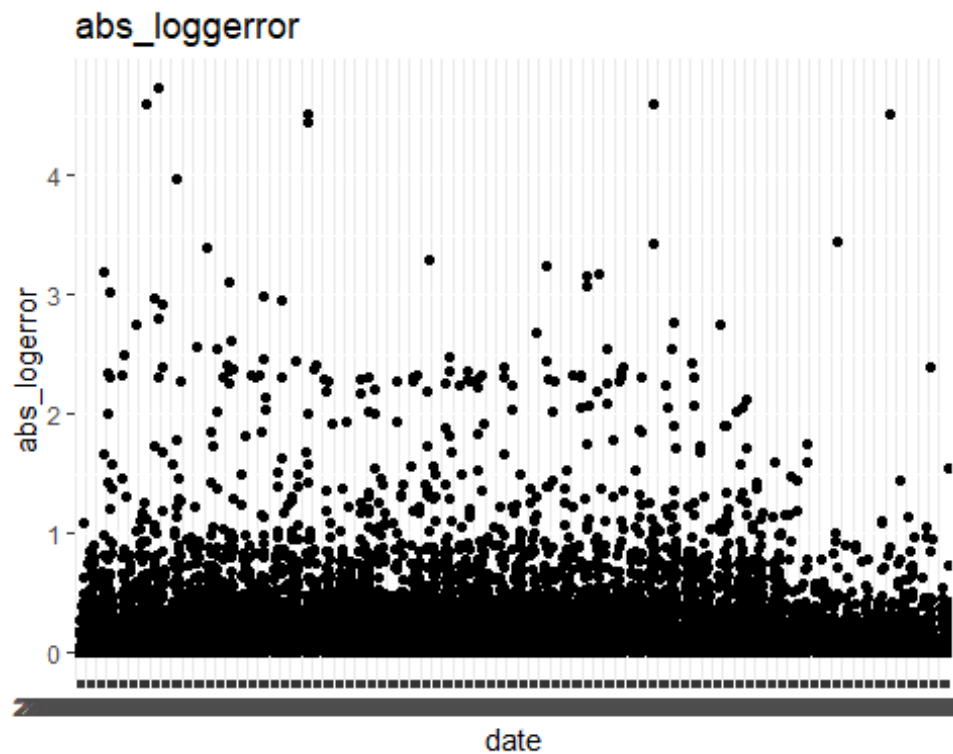


histogram of logerror binned



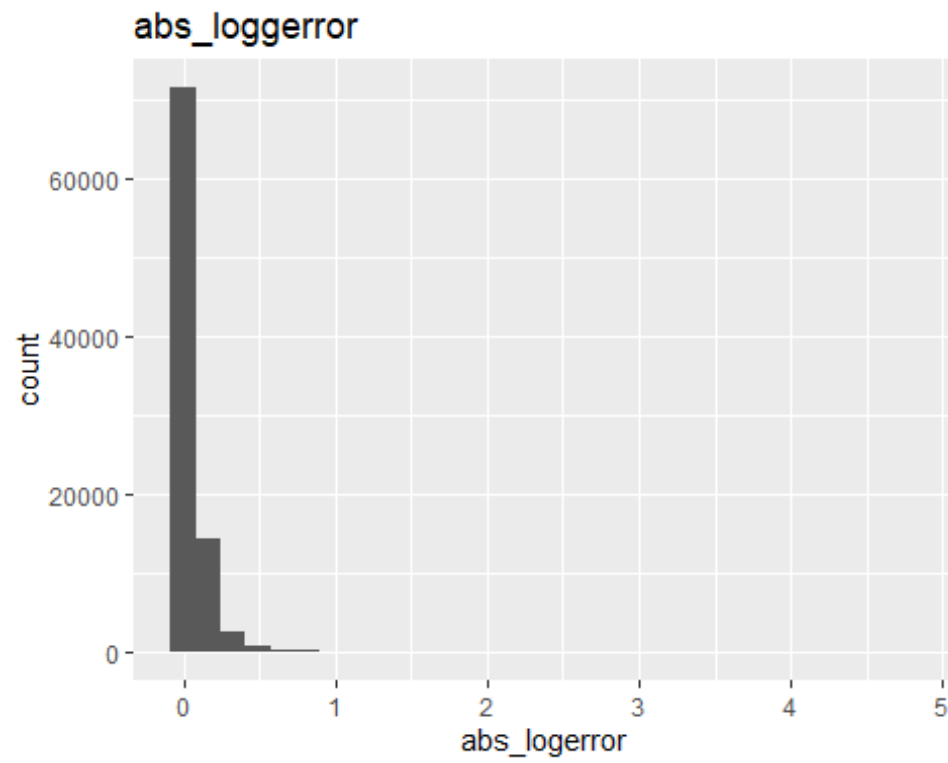
##	parcelid	logerror	date	year_month	abs_logerror
## 1	11016594	0.0276	2016-01-01	2016-01-01	0.0276
## 2	14366692	-0.1684	2016-01-01	2016-01-01	0.1684
## 3	12098116	-0.0040	2016-01-01	2016-01-01	0.0040
## 4	12643413	0.0218	2016-01-02	2016-01-01	0.0218
## 5	14432541	-0.0050	2016-01-02	2016-01-01	0.0050
## 6	11509835	-0.2705	2016-01-02	2016-01-01	0.2705

Scatter plot of abs_loggerror

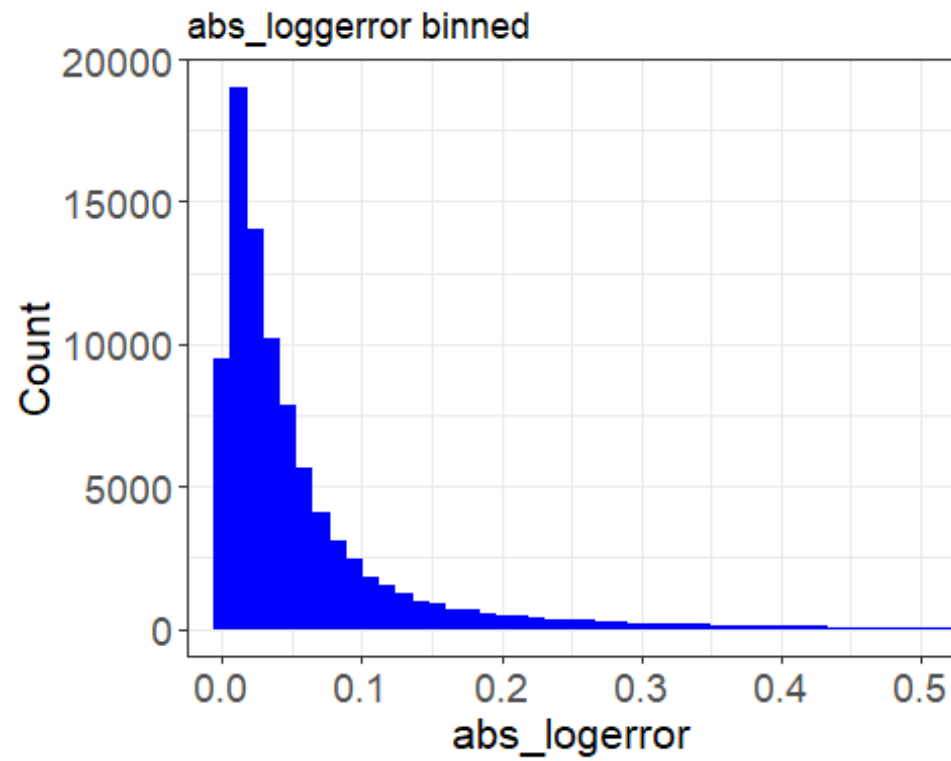


##	parcelid	logerror	date	year_month	abs_loggerror
## 1	11016594	0.0276	2016-01-01	2016-01-01	0.0276
## 2	14366692	-0.1684	2016-01-01	2016-01-01	0.1684
## 3	12098116	-0.0040	2016-01-01	2016-01-01	0.0040
## 4	12643413	0.0218	2016-01-02	2016-01-01	0.0218
## 5	14432541	-0.0050	2016-01-02	2016-01-01	0.0050
## 6	11509835	-0.2705	2016-01-02	2016-01-01	0.2705

Histogram of abs_logerror

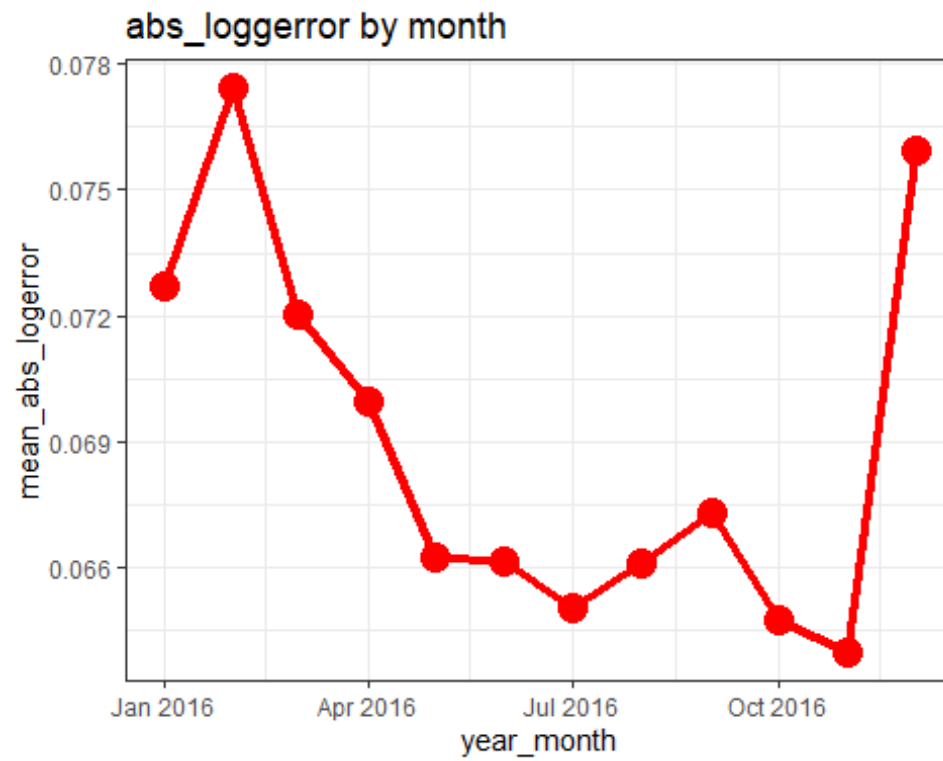


Histogram of abs_loggerror binned

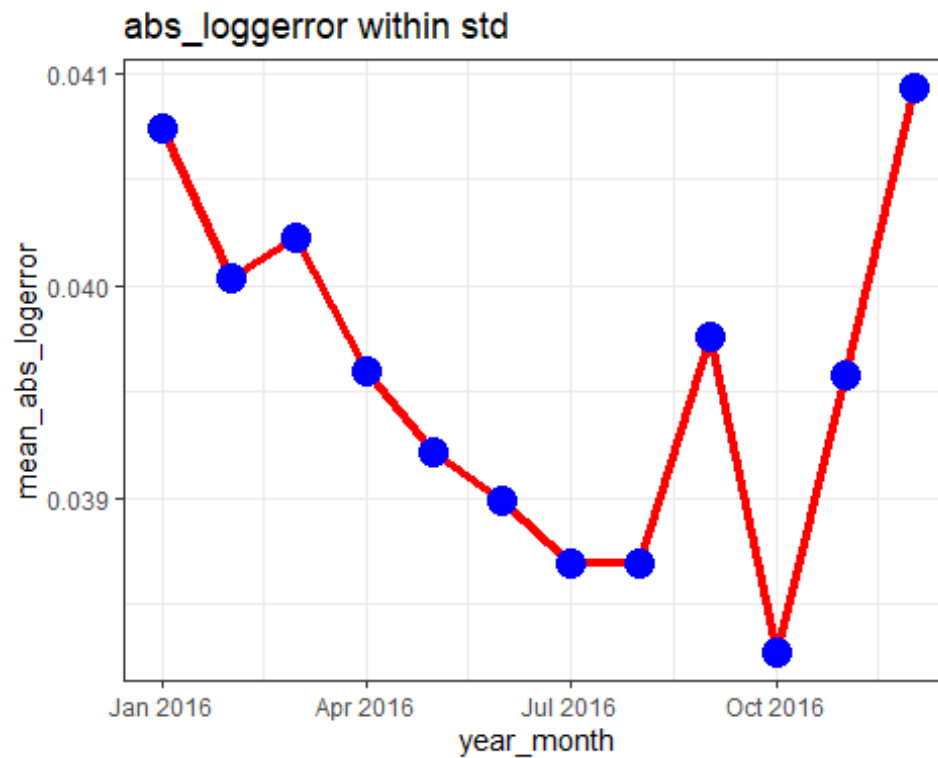


```
## geom_bar: na.rm = FALSE
## stat_bin: binwidth = NULL, bins = NULL, na.rm = FALSE, pad = FALSE
## position_stack
```

graph of abs_logerror groupped by month of year



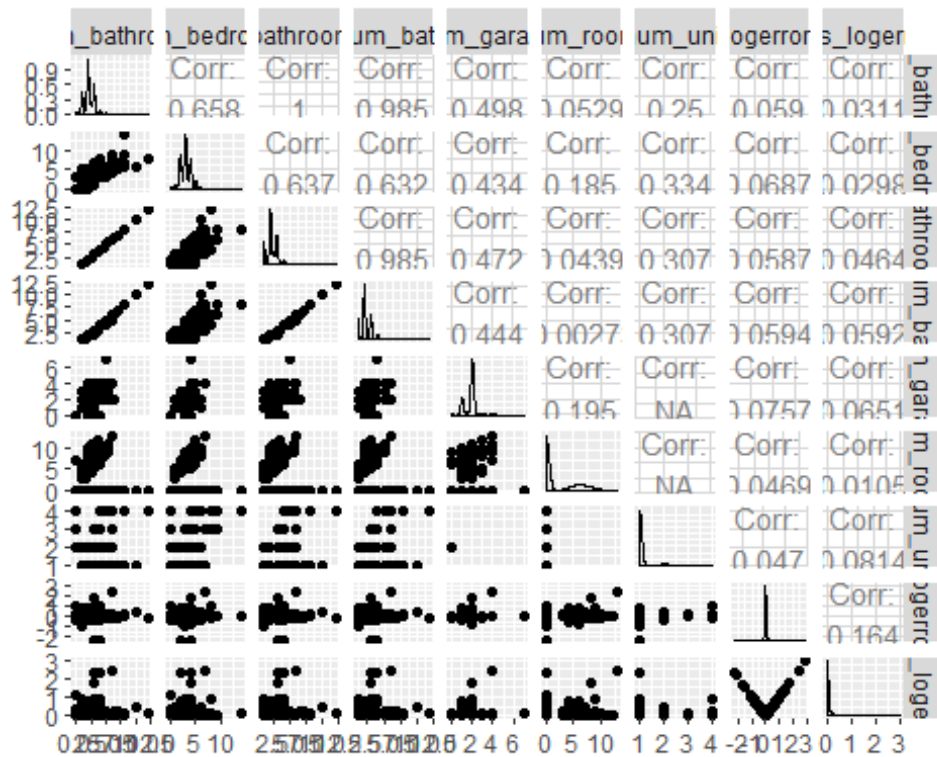
graph of abs_logerror within standard deviation



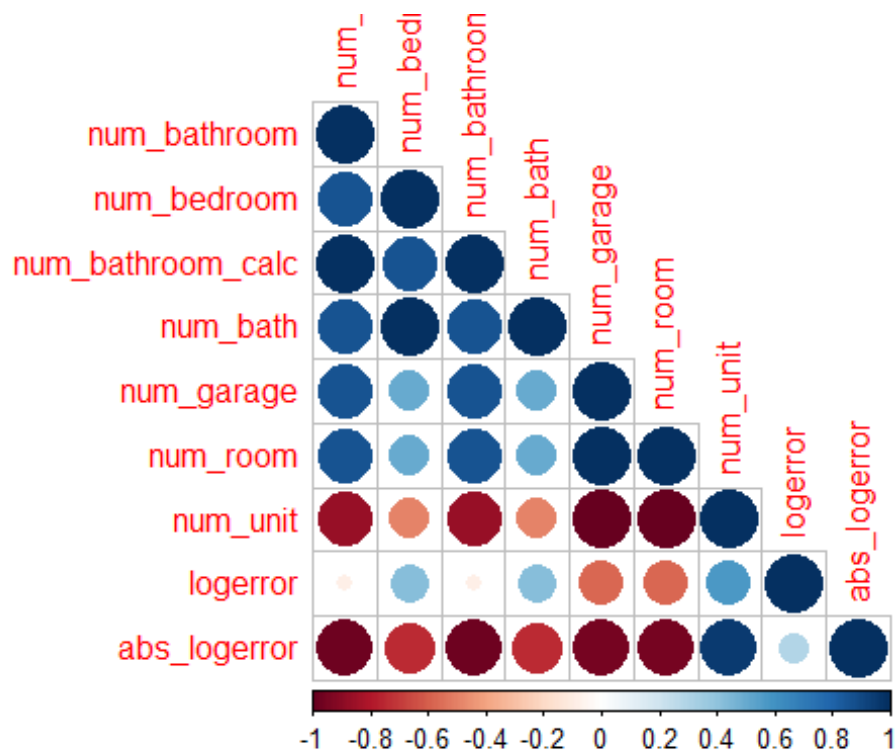
Correlations

##	num_bathroom	num_bedroom	num_bathroom_calc	num_bath	num_garage	num_room
## 1	NA	NA	NA	NA	NA	NA
## 2	NA	NA	NA	NA	NA	NA
## 3	3	2	3	3	NA	0
## 4	NA	NA	NA	NA	NA	NA
## 5	NA	NA	NA	NA	NA	NA
## 6	4	4	4	4	NA	0
##	num_unit	logerror	abs_logerror			
## 1	NA	0.0276	0.0276			
## 2	NA	-0.1684	0.1684			
## 3	1	-0.0040	0.0040			
## 4	NA	0.0218	0.0218			
## 5	NA	-0.0050	0.0050			
## 6	1	-0.2705	0.2705			

correlation using ggpairs using a subset of data(~10%)

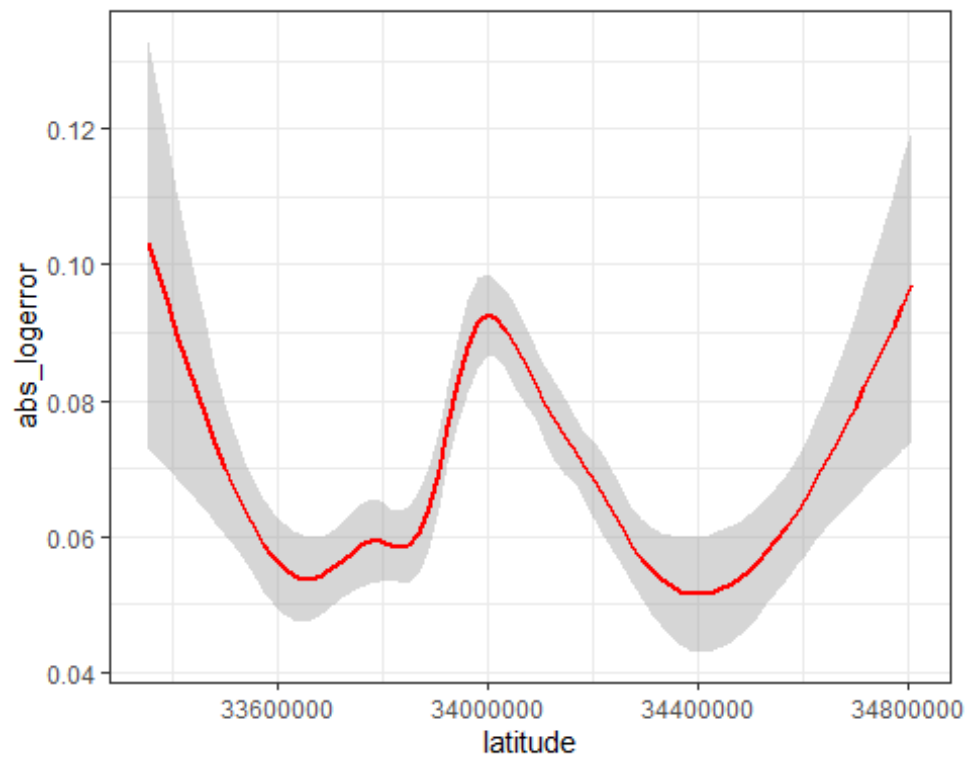
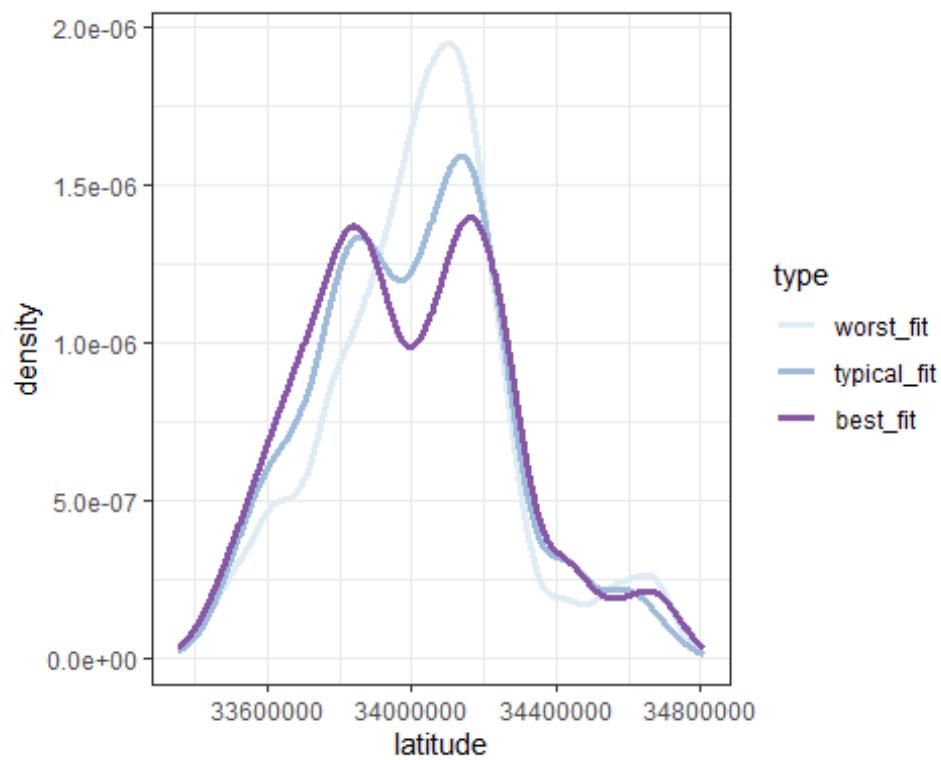


correlation using corplot

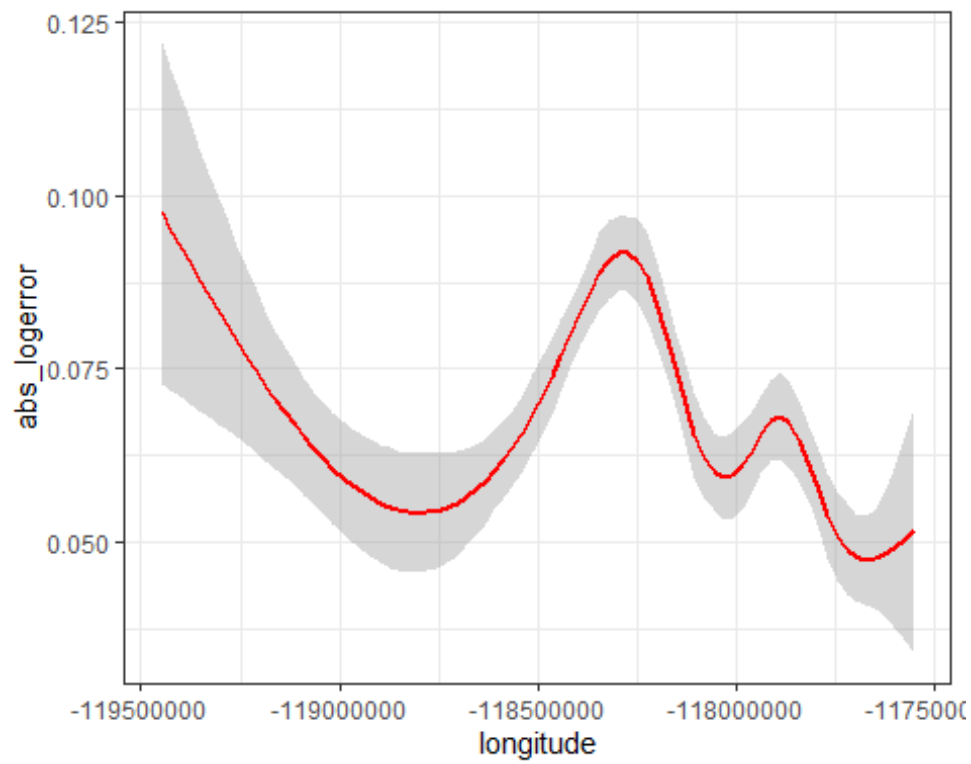
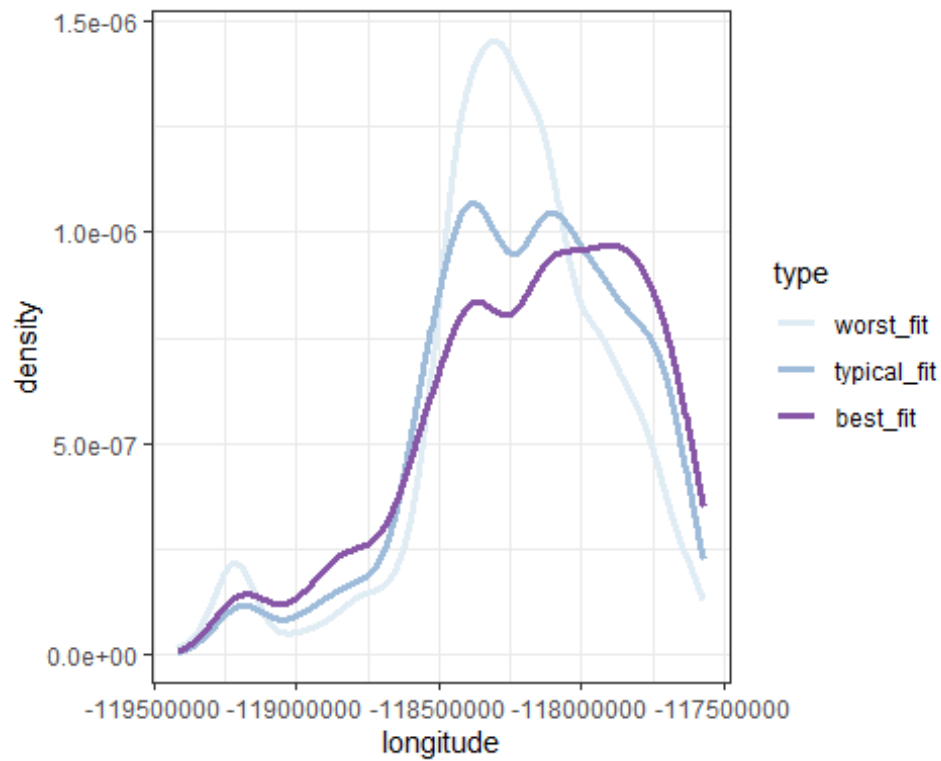


##	parcelid	logerror	date	year_month	abs_logerror	percentile
## 1	11016594	0.0276	2016-01-01	2016-01-01	0.0276	3
## 2	14366692	-0.1684	2016-01-01	2016-01-01	0.1684	5
## 3	12098116	-0.0040	2016-01-01	2016-01-01	0.0040	1
## 4	12643413	0.0218	2016-01-02	2016-01-01	0.0218	3
## 5	14432541	-0.0050	2016-01-02	2016-01-01	0.0050	1
## 6	11509835	-0.2705	2016-01-02	2016-01-01	0.2705	5

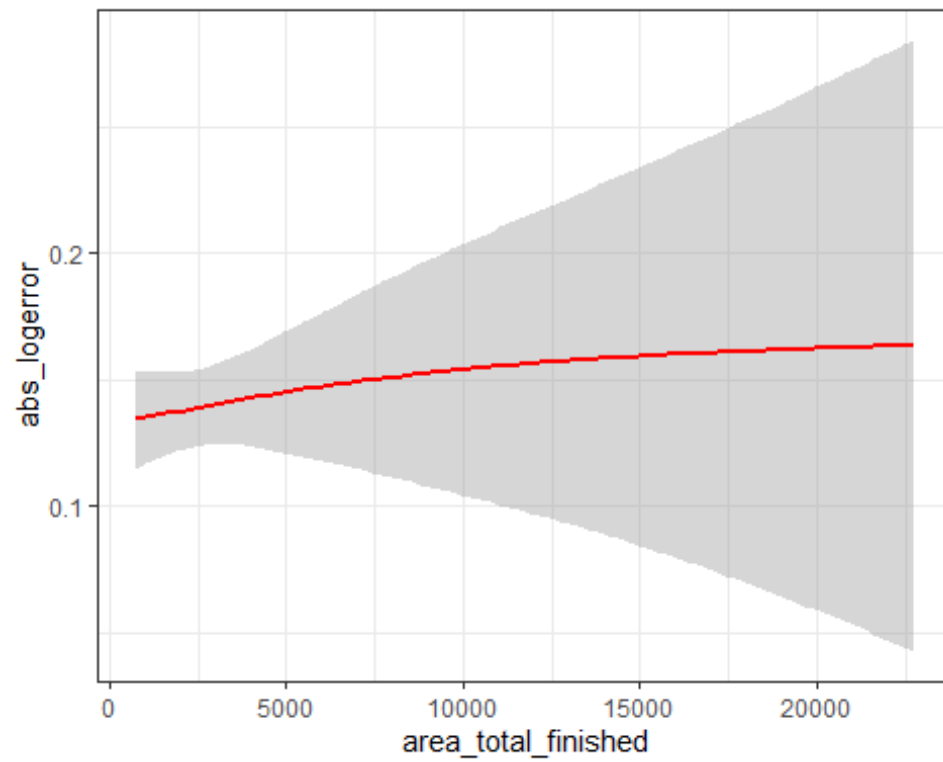
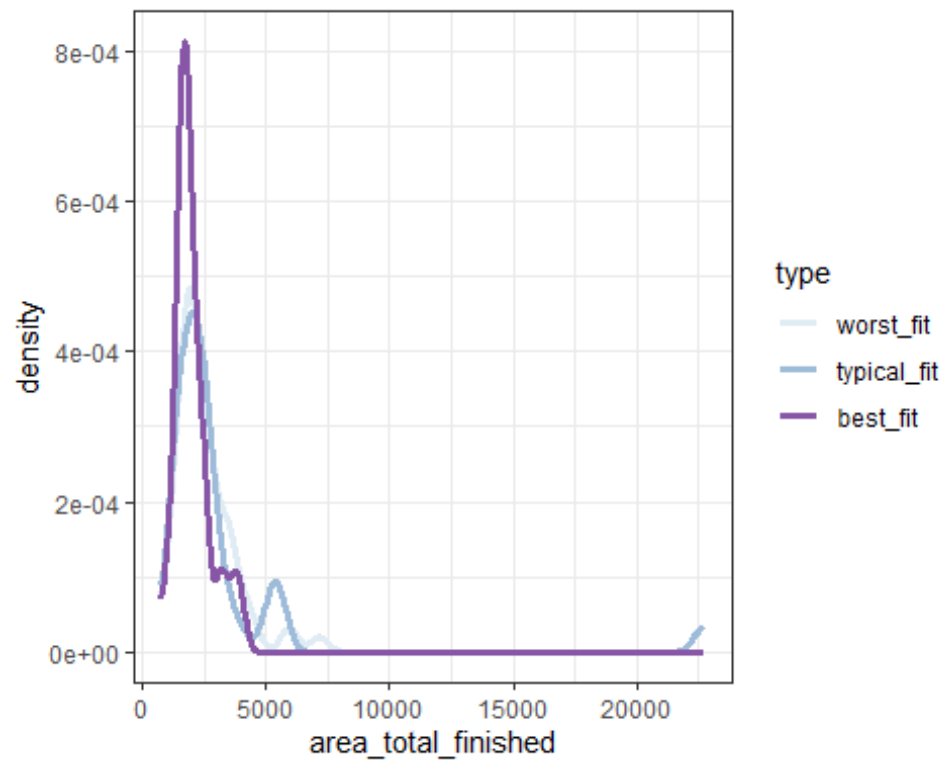
Density plot shows density of worst predicitions is lower in lower latitudes , but higher in median latitude and lower around and above 34400000.



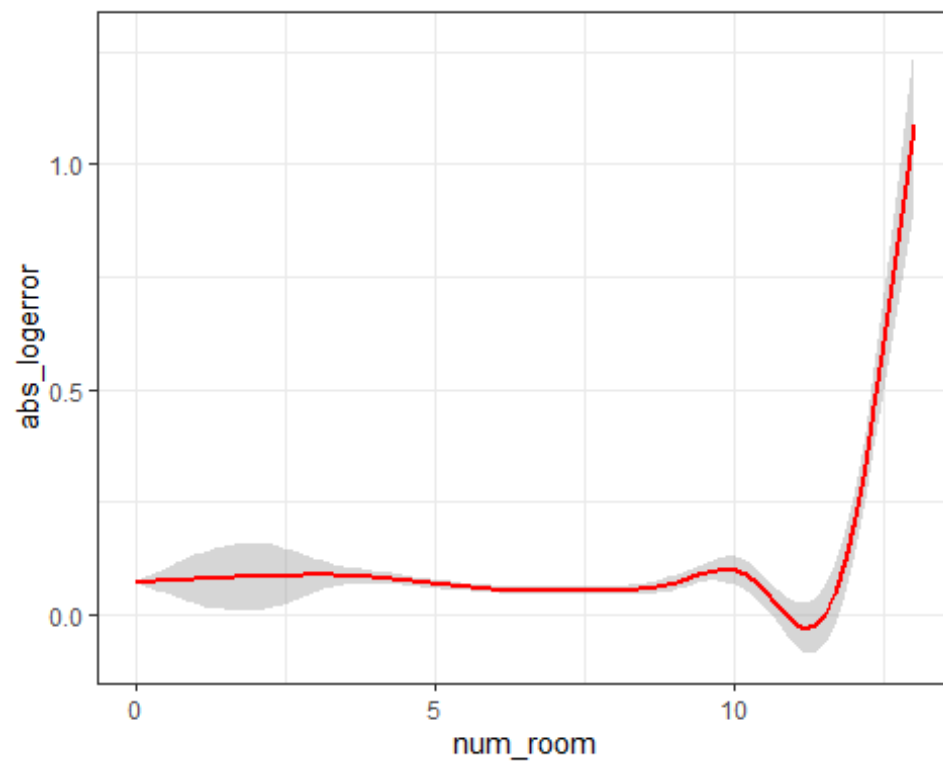
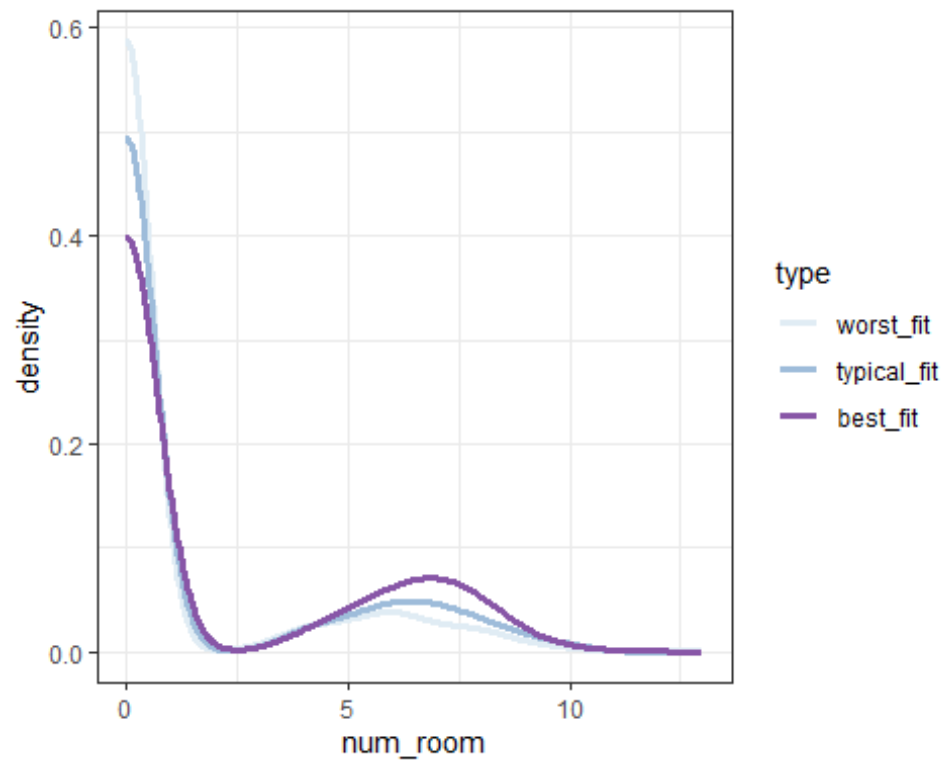
density of longitude. Worst predictions are in the -118500000 and -118000000



area_total_finished

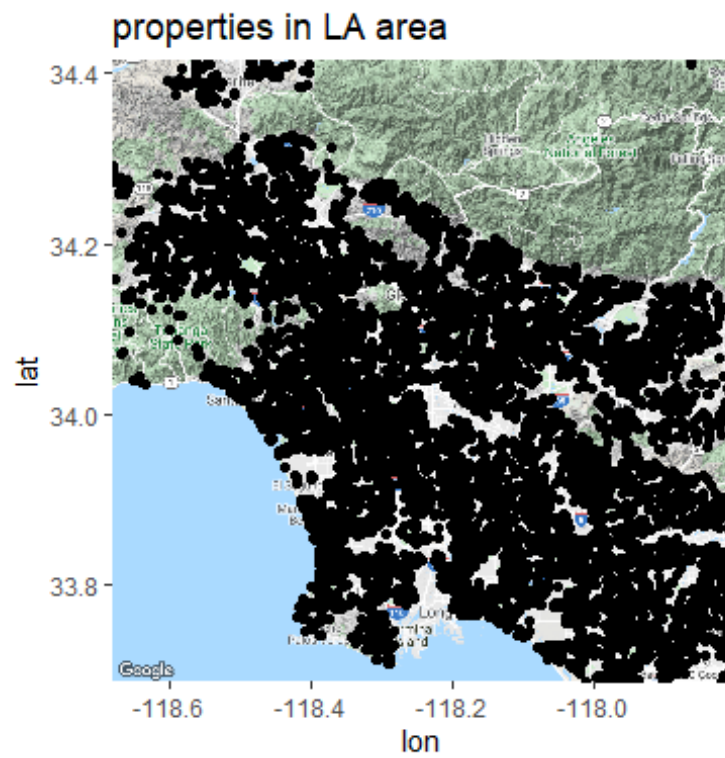


num_room

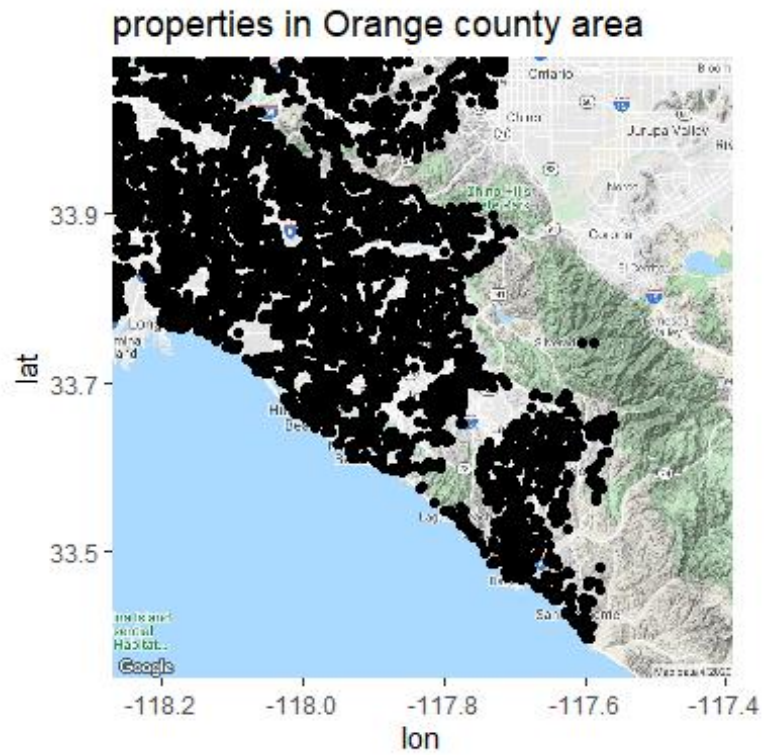


maps based on a sampling of 10000 properties

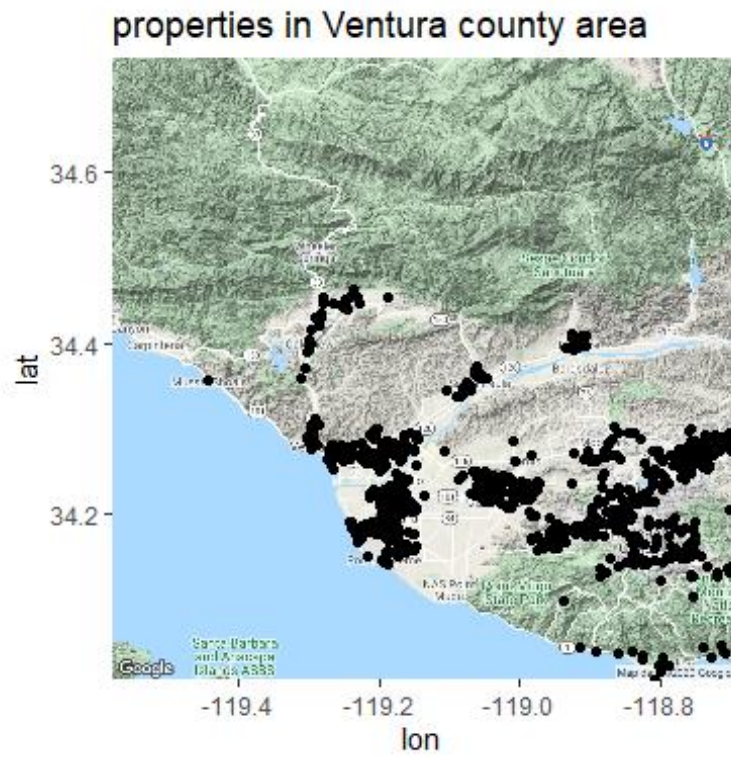
Map of Los Angeles county



Map of Orange County



Map of Ventura County



Absolute log error based on location

