Challenge 3   Name: Eagle Xuhui Ying   WFU ID: 06648543

1. Demonstrated deep understanding of the problem statement, what does it mean to City of Boston?

Suppose I recently joined Boston Consulting Group to clean up a mess left by a previous consultant for the city of Boston. The City of Boston is considering awarding our firm a large consulting contract if I can beat an RMSE of predictions on property tax assessments of $57854. To do this, I need to analyze and build some models to assess and predict the av_total (assessed value) of properties in the greater Boston area. I have to train and compare a Linear Regression, Random Forest, and XGBoost model.

1a. KEY INSIGHTS into Factors Influencing AV Total

(1) The distribution of av_total is right-skewed. The median of av_total is 418,700, the mean of av_total is 448,563.6

(2) The higher Parcel's land area is, the higher assessed housing price will be.

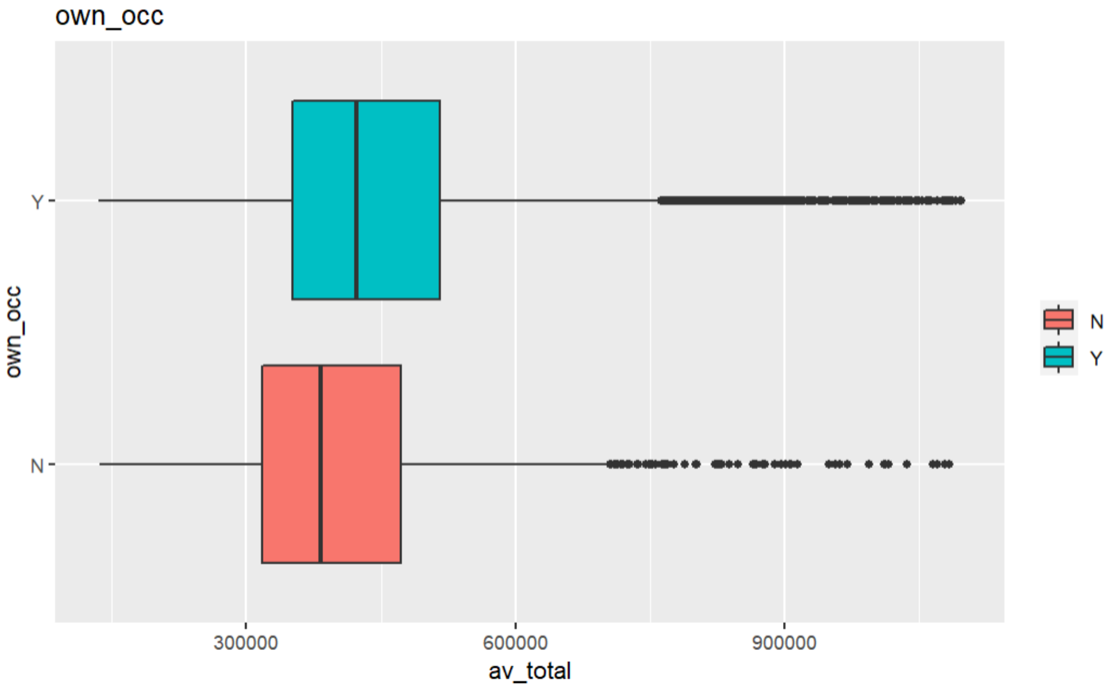
(3) The later houses are built, the lower assessed housing price will be.

(4) The assessed housing price is positive correlated with number of levels in the structure located on the parcel, total number of rooms, total number of bedrooms, total number of full baths, total number of half baths, and total number of fireplaces in the structure.

(5) The assessed housing price in Jamica Plain (Postcode: 2130) is higher than other four regions in Massachusetts.

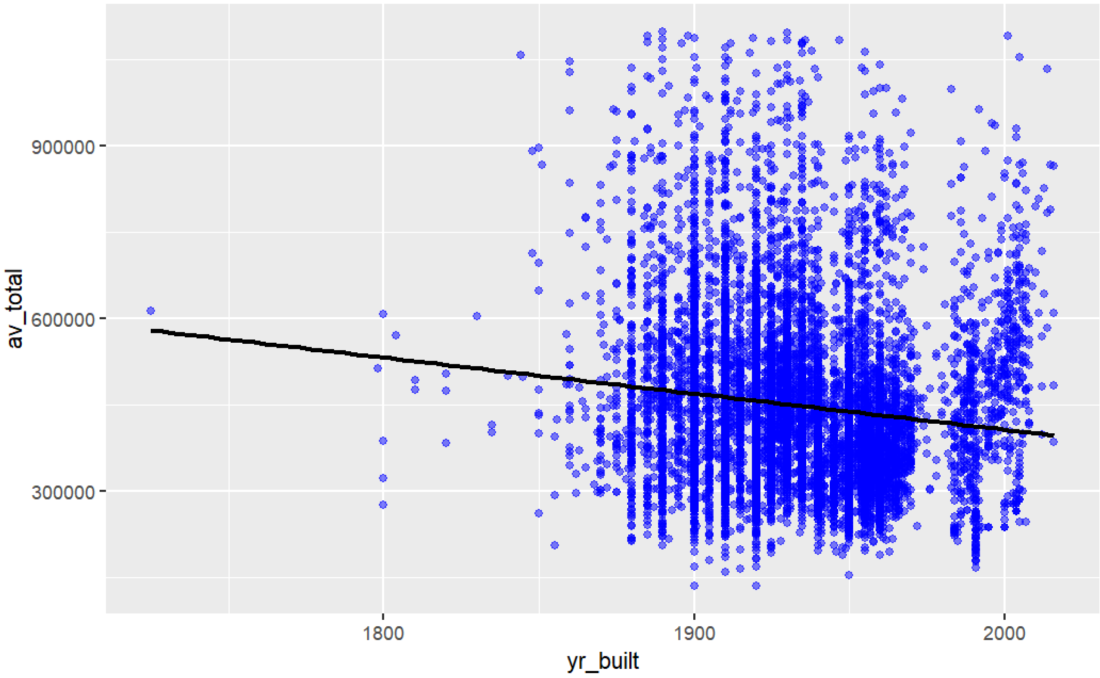
2. Answers key business questions, assertions, and beliefs with data.

a. The City of Boston believes that owner-occupied homes have a higher assessed value



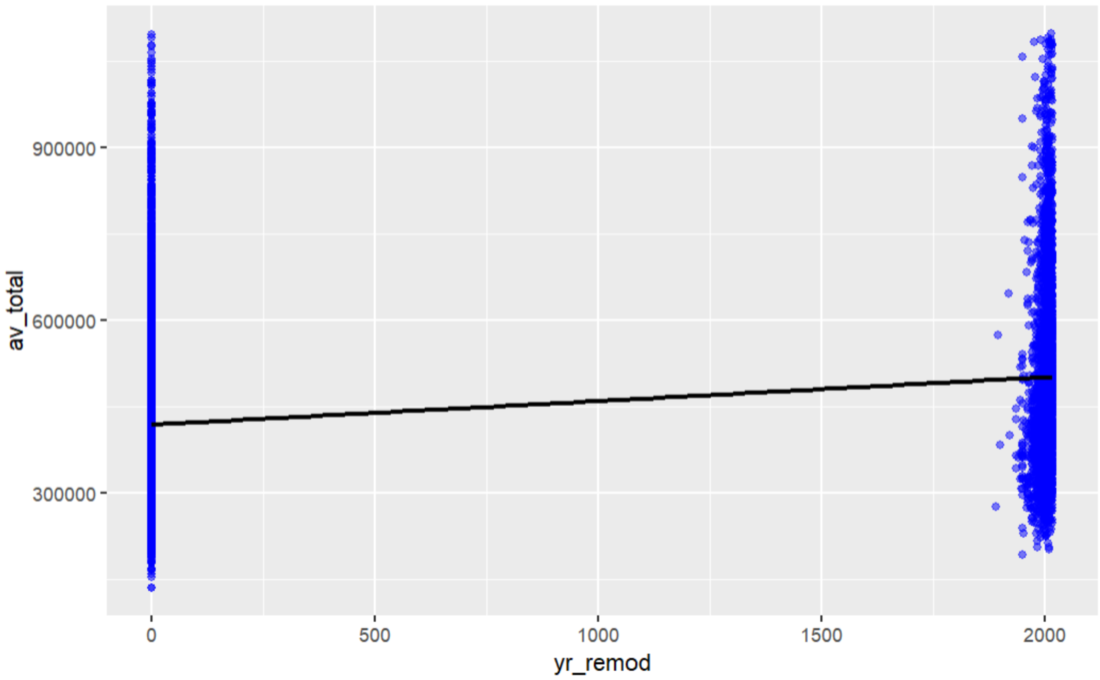
Yes. According to the boxplot, the median assessed value of owner-occupied homes is higher than homes without owners.

b. homes built in the 1990s, tend to have higher home values.



No. According to the scatterplot, the earlier home is built, the higher assessed value they tend to have. Thus, homes built in the 1990s tend to have lower home values.

c. homes that have been recently remodeled tend to have higher home values.



Yes. According to the scatterplot, homes remodeled around 2000 tend to have higher assessed value than homes without remodeling.

3. Explanation, definition, and justification of evaluation metrics

(1) Linear Regression: Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes

the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. Linear mode power supplies offer many advantages such as a simple design and overall low cost while also having disadvantages like high heat loss and varied, low efficiency levels.
(2) Random Forests: The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It can perform both regression and classification tasks, while a large number of trees can make the algorithm too slow and ineffective for real-time predictions.
(3) XGBoost: XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. It is Effective with large data sets. However, it can over-fit the data, especially if the trees are too deep with noisy data.

4. Actionable recommendations, using model output and analysis insight.
(1) If buyers wants to spend less money to buy a house in Boston, they can consider recently built houses with low Parcel's land area.
(2) If owners want to sell their houses in higher prices, they are increase the number of levels in the structure located on the parcel, total number of rooms, total number of bedrooms, total number of full baths, total number of half baths, and total number of fireplaces in the structure.
(3) The Boston Consulting Group should focus on the Jamica Plain (Postcode: 2130), where housing prices is obviously higher than other four regions in Massachusetts. They should analyze the reason behind it, and other potential predictors which may influence the housing price in this area.

5. Clear methodology (what steps are you taking to prepare and evaluate the models)
(1) Data partitioning
- Split the data into 70/30 train/test split using random sampling

(2) Data preprocessing
- Formula

i. av_total ~ land_sf + yr_built + living_area + num_floors + r_total_rms + r_bdrms + r_full_bth + r_half_bth + r_kitch + r_fplace + own_occ + r_bldg_styl + r_roof_typ + r_ext_fin + r_bth_style + r_kitch_style + r_heat_typ + r_ac + r_ext_cnd + r_ovrall_cnd + r_int_cnd + r_int_fin + r_view + city_state
- Numeric Predictor Pre-Processing

i. Replaced missing numeric variables with median
ii. Use an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time
- Categorical Predictor Pre-Processing

i. Replaced missing categorical variables with “unknown”

ii. Dummy encoded categories with 1s and 0s

(3) Model specification

- Model 1: Linear Regression
- Model 2: Random Forest
- Model 3: XGBoost

trees <int>	tree_depth <int>	learn_rate <dbl>	.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>
1029	1	0.086337954	rmse	standard	62218.17	5	610.8497
1405	6	0.001243759	rmse	standard	103935.39	5	1093.7883
1730	8	0.204655494	rmse	standard	54410.83	5	950.4221
561	10	0.018774356	rmse	standard	53586.78	5	1097.7259
362	13	0.005236247	rmse	standard	94773.18	5	1175.1681
457	9	0.316047405	rmse	standard	56581.54	5	798.0225
1995	13	0.051233099	rmse	standard	54427.52	5	976.0895
826	15	0.028471219	rmse	standard	54667.28	5	1084.9996
944	4	0.023913257	rmse	standard	52507.42	5	726.6163
1994	6	0.021032715	rmse	standard	52027.90	5	974.7142

trees <int>	tree_depth <int>	learn_rate <dbl>	.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>
1063	15	0.166371298	rmse	standard	55764.60	5	1065.5478
1061	8	0.027816190	rmse	standard	53163.82	5	982.5086
1933	2	0.019987968	rmse	standard	54974.39	5	576.8134
1383	7	0.023673087	rmse	standard	52477.77	5	898.3286
155	1	0.305140526	rmse	standard	63297.99	5	581.3941

The tuning process of the XGB model generates 10 iterations in this case and the best one (with the lowest mean of 52030) is iteration 5 with a tree number of 1994, tree depth of 6, and learn_rate of 0.021.

6. Clear model metrics and evaluation of 3 or more models

d. RMSE: The Root mean square erro (RMSE) of an estimator of a population parameter is the square root of the mean square error (MSE). The mean square error is defined as the expected value of the square of the difference between the estimator and the parameter. It is the sum of variance and squared Bias.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

e. R-square: The R-squared value is the amount of variance explained by your model. It is a measure of how well your model fits your data. As a matter of fact, the higher it is, the better is your model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

f. MAE: Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. The closer MAE is to 0, the more accurate the model is.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

g. R2, RMSE, MAE

Model	Partition	RMSE	R-square	MAE
Linear Regression	train	65511.34	0.7969257	45396.35
Random Forest	train	48501.91	0.9138376	36824.91
XGBoost	train	23351.54	0.975636	17292.9
Model	Partition	RMSE	R-square	MAE
Linear Regression	test	62198.37	0.8249286	44343.42
Random Forest	test	60334.93	0.8496907	44213.79
XGBoost	test	52178.72	0.8708404	37418.88

According to the above table, the RMSE of XGBoost model is lower among the three tested models. There, in this case, XGBoost model is better than Linear Regression and Random Forest.

7. Data dictionary

Variable	Definition	Keep
pid	Unique 10-digit parcel number	ignore
zipcode	Zip code of parcel	ignore
own_occ	One-character code indicating if owner receives residential exemption as an owner-occupied property	keep
av_total	Assessed value for property i.e. what you are predicting	target
land_sf	Parcel's land area in square feet (legal area)	keep
yr_built	Year property was built	keep
yr_remod	Year property was last remodeled	keep
living_area	Living area square footage of the property	keep
num_floors	# of levels in the structure located on the parcel	keep
structure_class	Structural classification of commercial building	ignore
r_bldg_styl	Residential building style	keep
r_roof_typ	Structure roof type	keep
r_ext_fin	Structure exterior finish	keep
r_total_rms	Total number of rooms in the structure	keep
r_bdrms	Total number of bedrooms in the structure	keep
r_full_bth	Total number of full baths in the structure	keep
r_half_bth	Total number of half baths in the structure	keep
r_bth_style	Residential bath style	keep
r_kitch	Total number of kitchens in the structure	keep
r_kitch_style	Residential kitchen style	keep
r_heat_typ	Structure heat type	keep
r_ac	Indicates if the structure has air conditioning (A/C)	keep
r_fplace	Total number of fireplaces in the structure	keep
r_ext_cnd	Residential exterior condition	keep
r_ovrall_cnd	Residential overall condition	ignore
r_int_cnd	Residential interior condition	keep

r_int_fin	Residential interior finish	keep
r_view	Residential view	keep
zip	ZIP CODE – should join to ZIPCODE	ignore
population	Population of people in the ZIP code	ignore
pop_density	People per square mile	ignore
median_income	Median Income of the residence of that zip code	ignore
city_state	City Name and State	keep

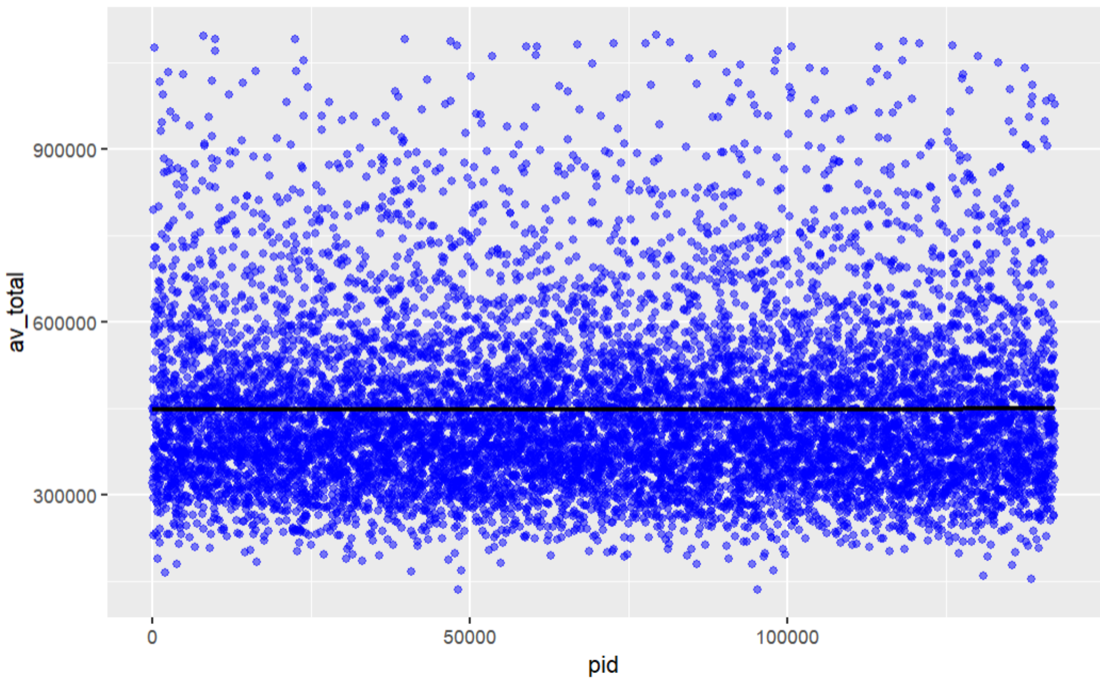
h. Addresses what variables are excluded: pid (they are different in every piece of record); structure_class (it only has one category); zipcode, r_ovrall_cnd, zip, population, pop_density, median_income (they are one-to-one corresponding with city_state)

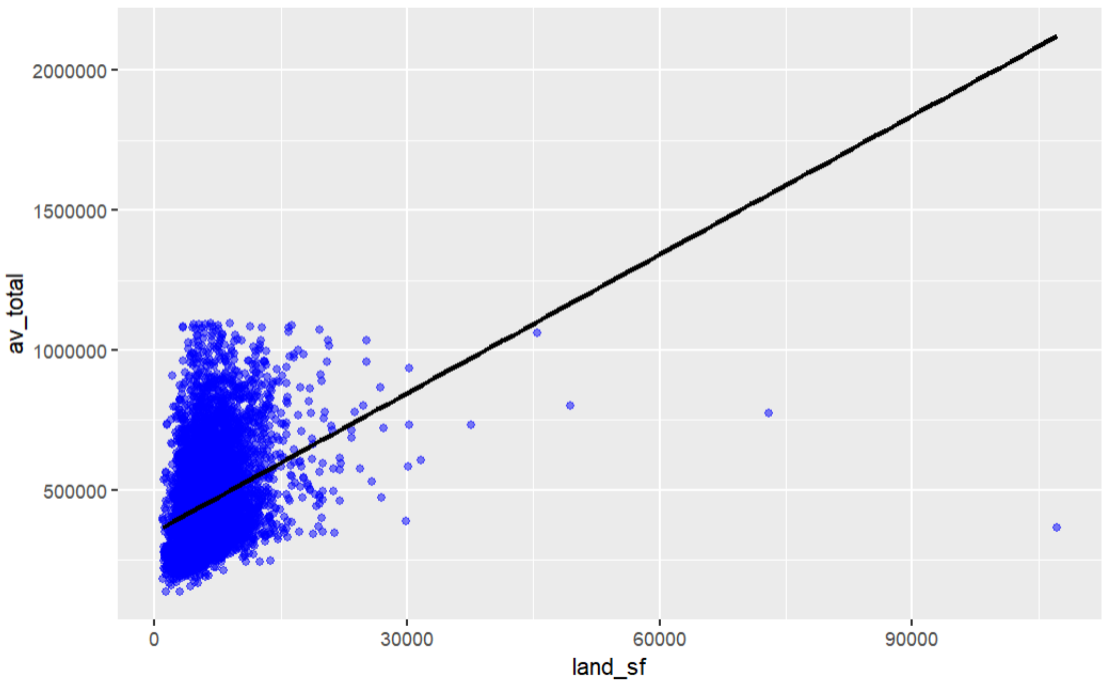
i. included, their role: land_sf, yr_built, living_area, num_floors, r_total_rms, r_bdrms, r_full_bth, r_half_bth, r_kitch, r_fplace, own_occ, r_bldg_styl, r_roof_typ, r_ext_fin, r_bth_style, r_kitch_style, r_heat_typ, r_ac, r_ext_cnd, r_ovrall_cnd, r_int_cnd, r_int_fin, r_view, city_state (They are used to predict the housing price in boston)

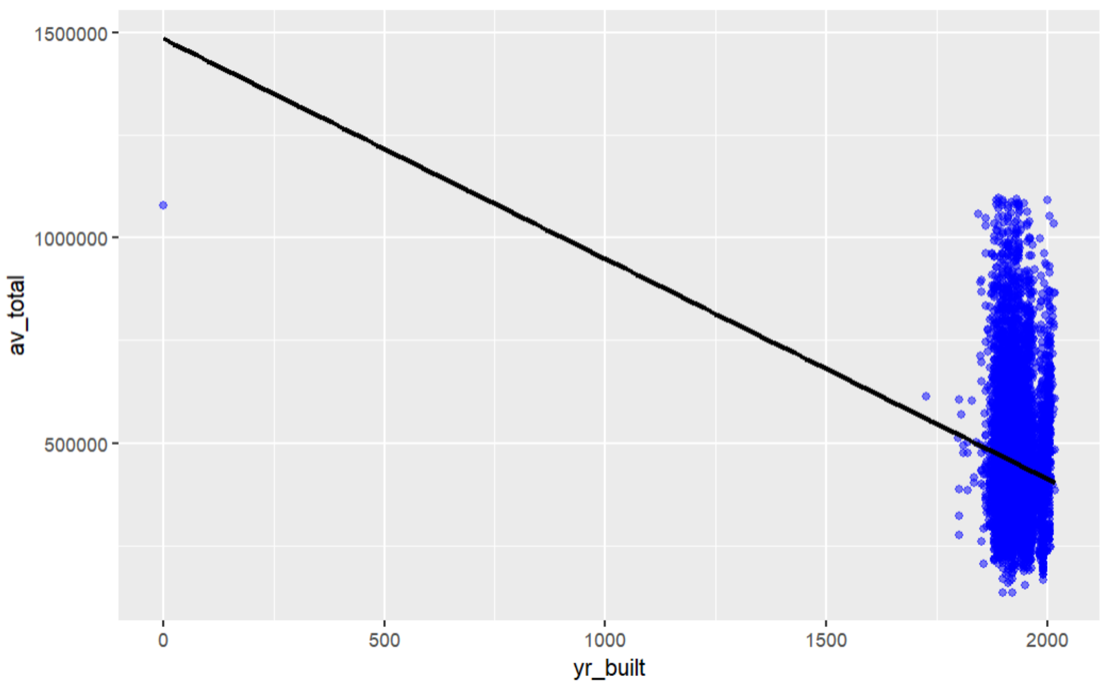
j. expected transformations: Since the distribution of av_total is right-skewed, there are more inexpensive houses than expensive ones. When modeling this type of outcome, a strong argument can be made that the price should be log-transformed. The advantages of this type of transformation are that no houses would be predicted with negative sale prices and that errors in predicting expensive houses will not have an undue influence on the model.

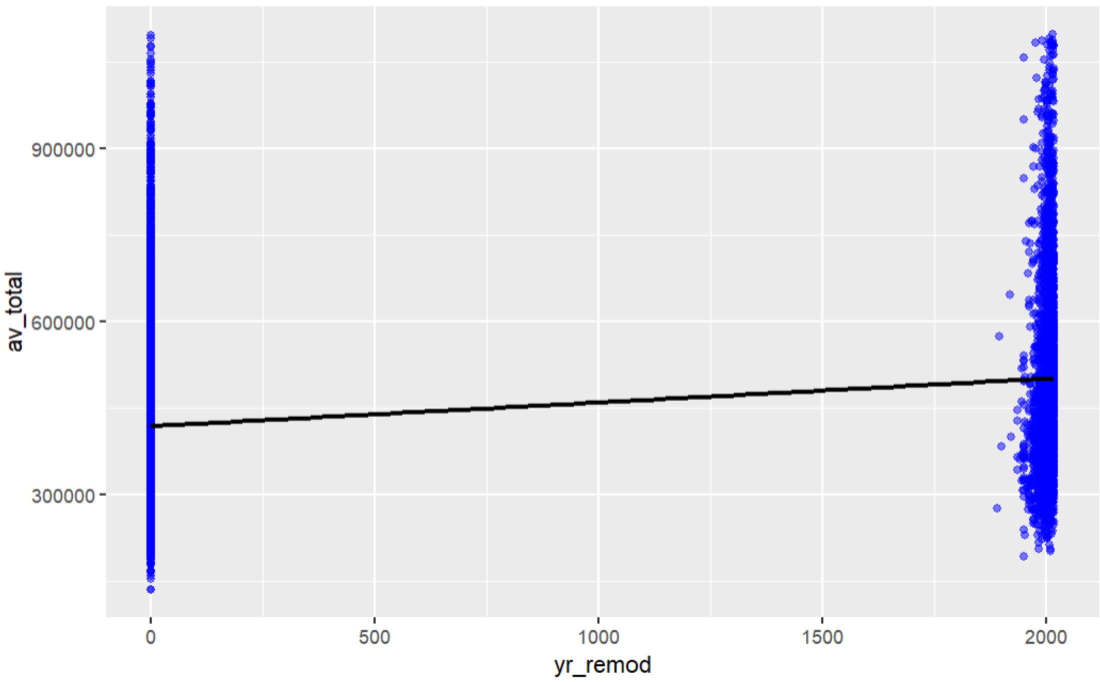
8. Any supporting exploratory data analysis (EDA)

(1) Numerical Variables: pid, land_sf, yr_built, yr_remod, living_area, num_floors, r_total_rms, r_bdrms, r_full_bth, r_half_bth, r_kitch, r_fplace, population, pop_density, median_income



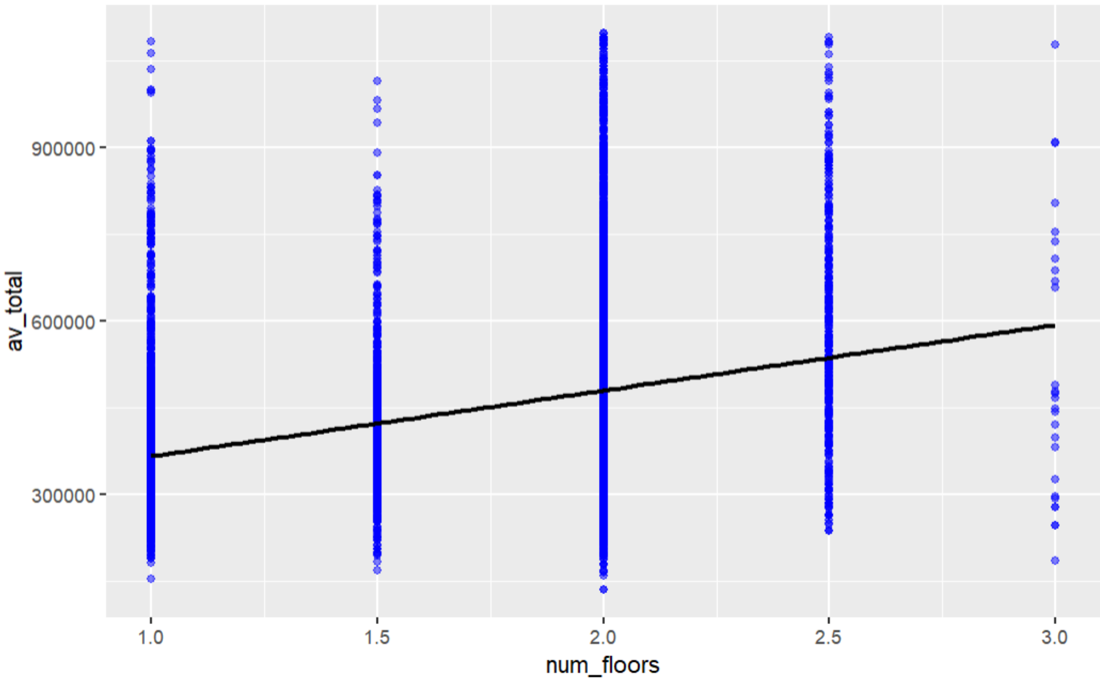

2000000
1500000
1000000
500000
av_total
0
30000
60000
90000
land_sf


1500000
1000000
500000
av_total
0
500
1000
1500
2000
yr_built

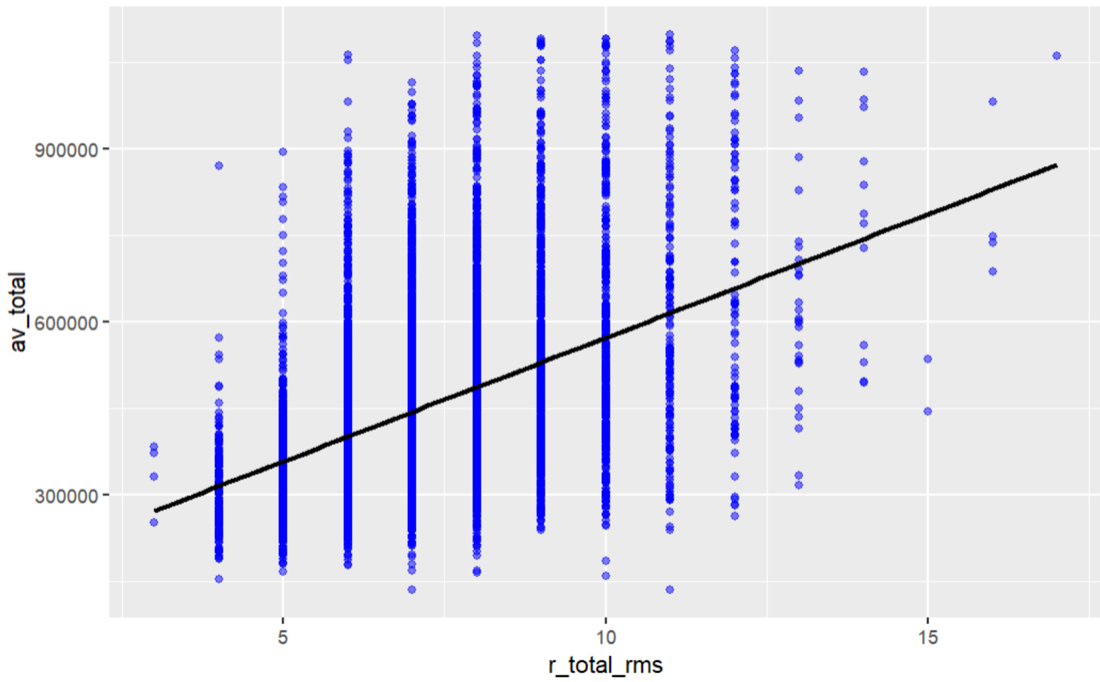


900000
600000
300000
av_total
0
500
1000
1500
2000
yr_remod

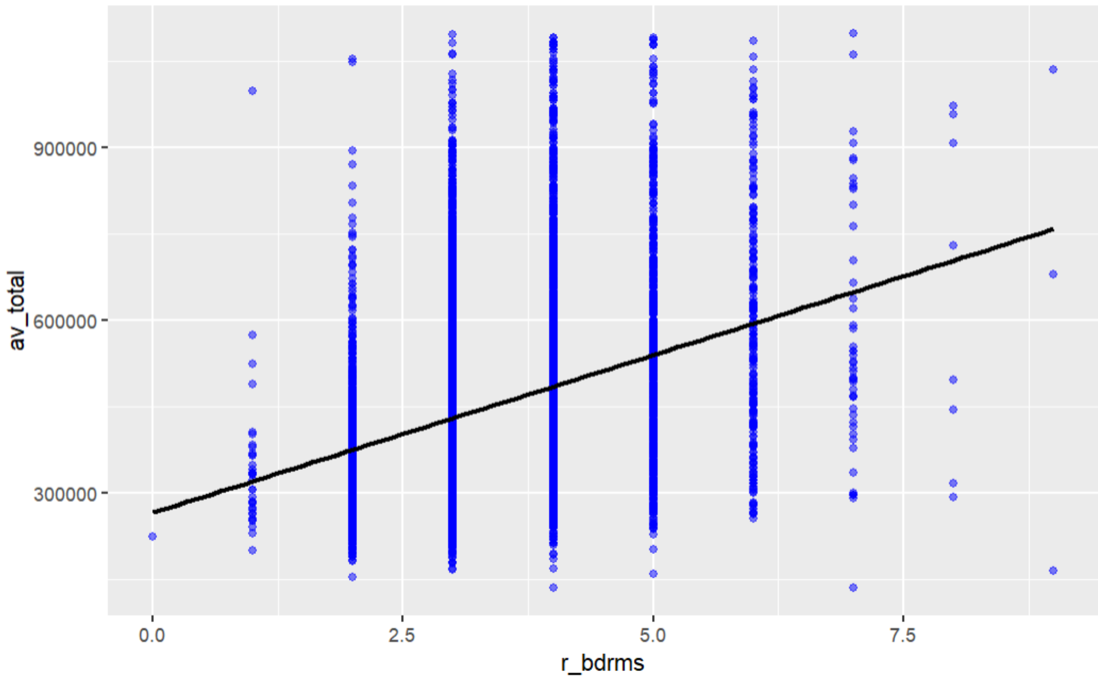
1600000
1200000
800000
400000
av_total
0
2500
5000
7500
living_area

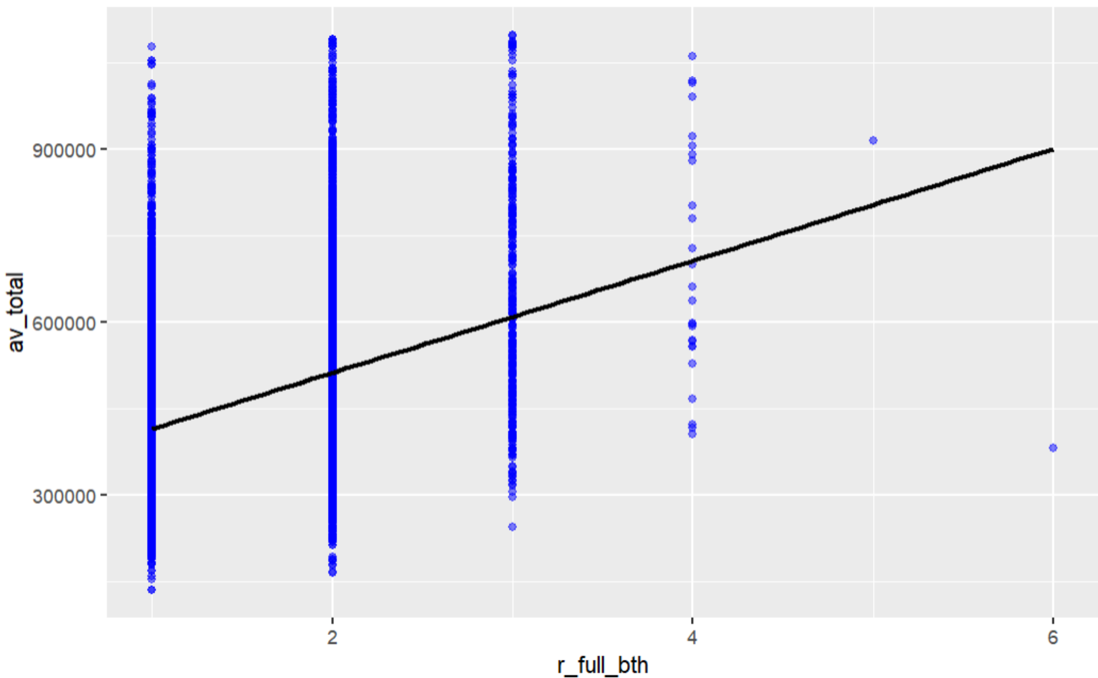


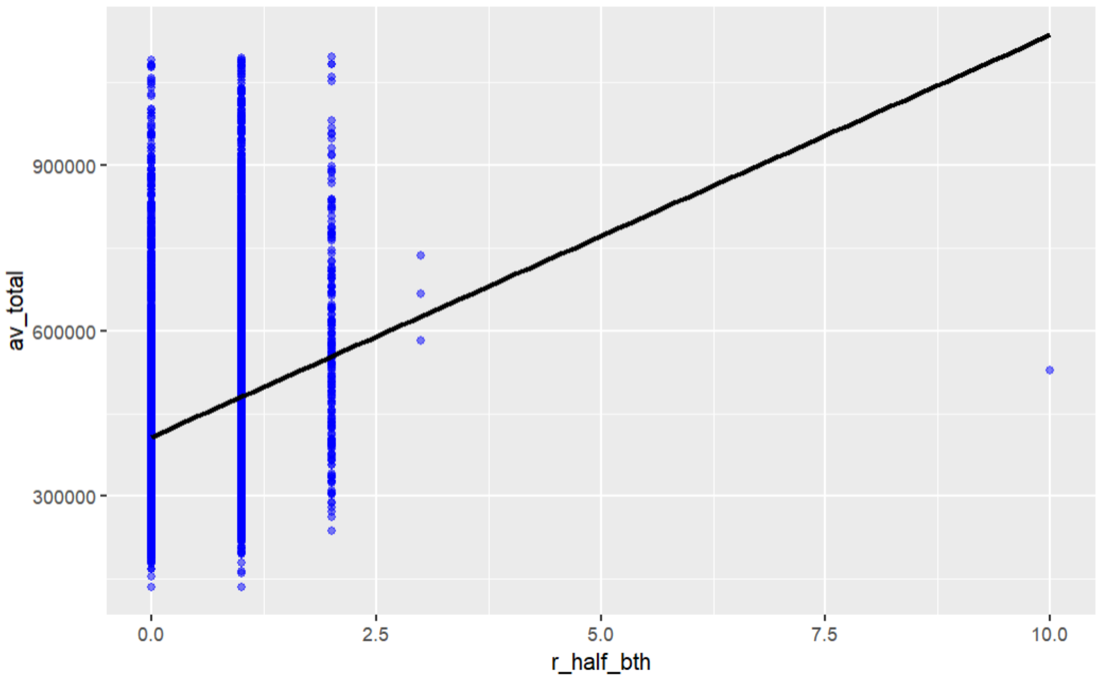
av_total
900000
600000
300000
1.0
1.5
2.0
2.5
3.0
num_floors

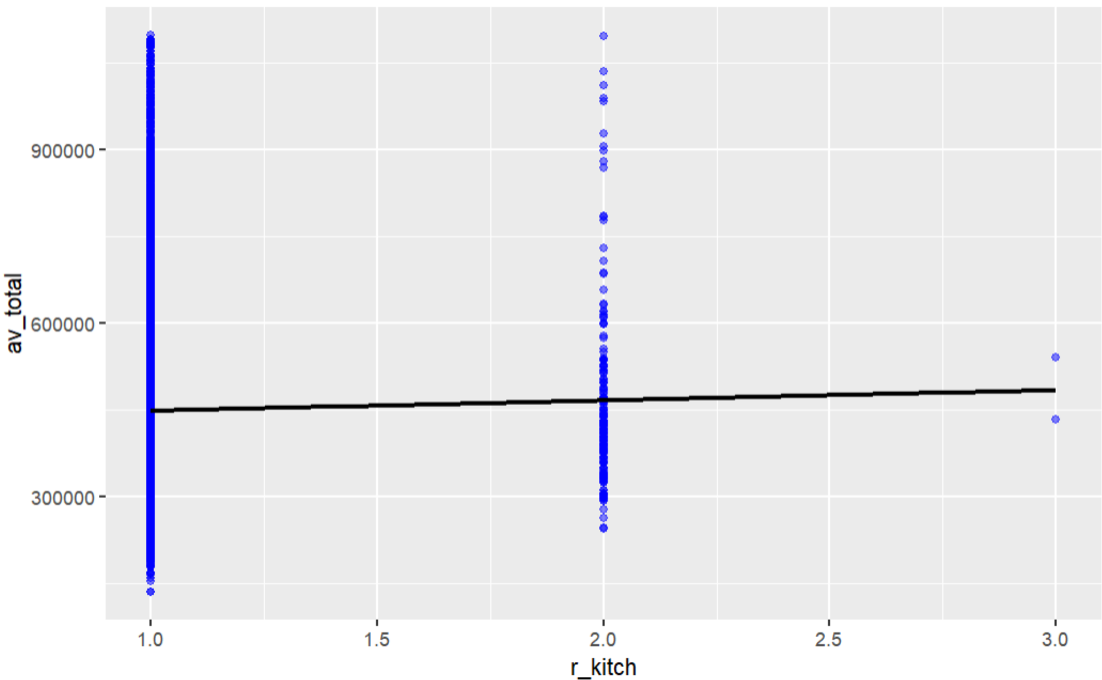


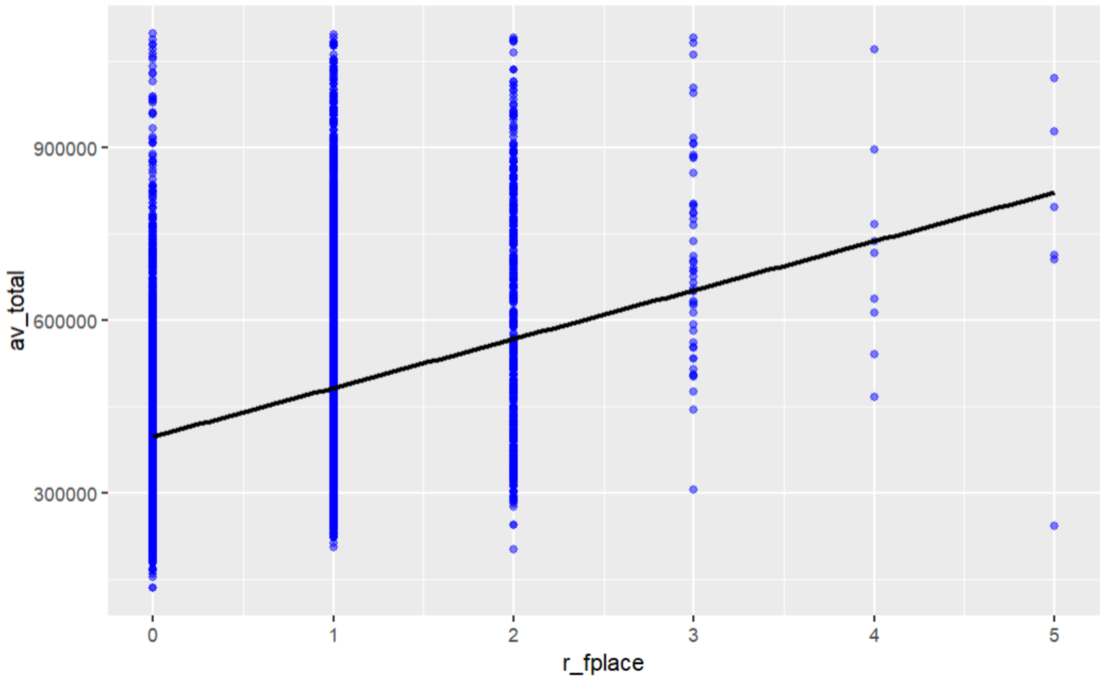
av_total
900000
600000
300000
5
10
15
r_total_rms

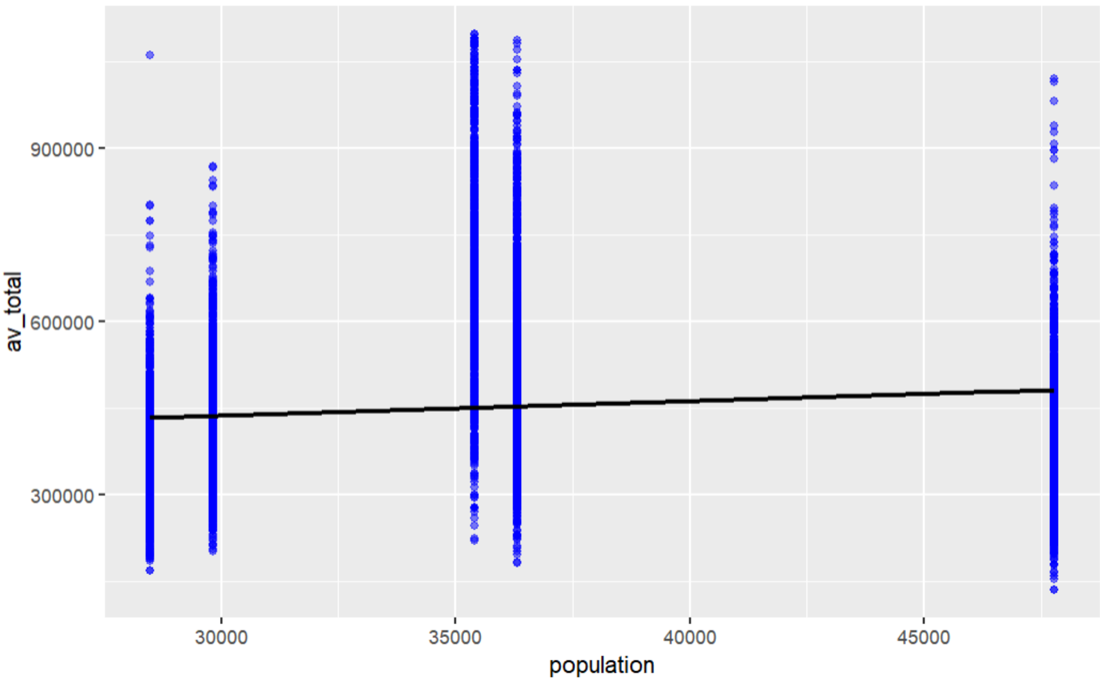

900000
600000
300000
av_total
0.0
2.5
5.0
7.5
r_bdrms

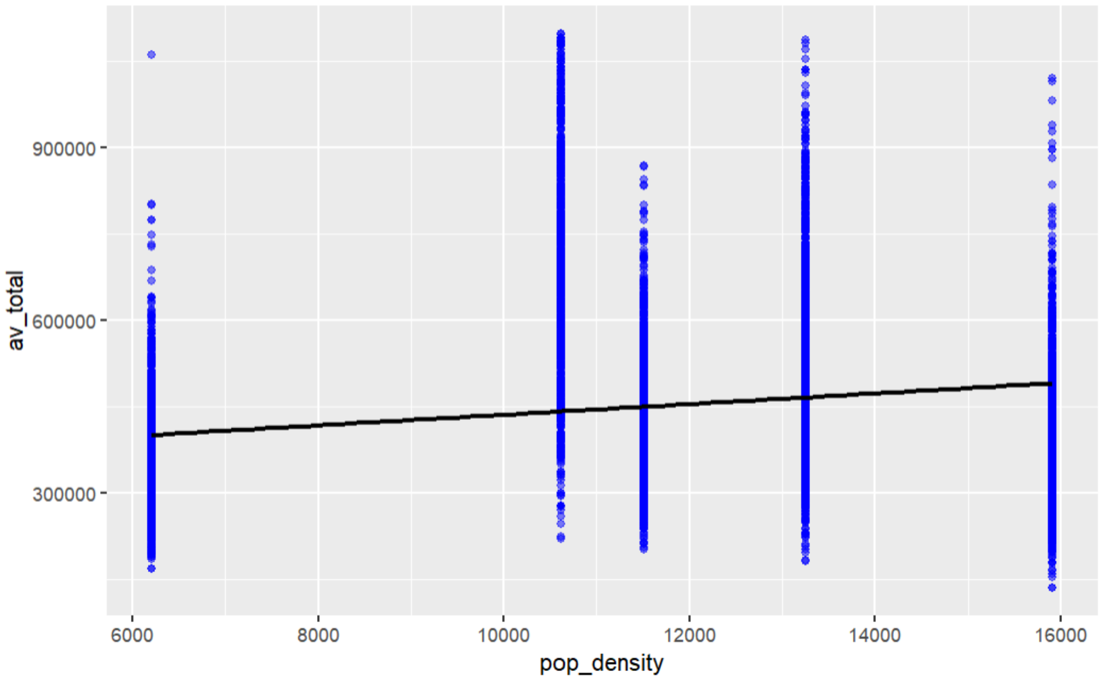

900000
600000
300000
av_total
2
4
6
r_full_bth

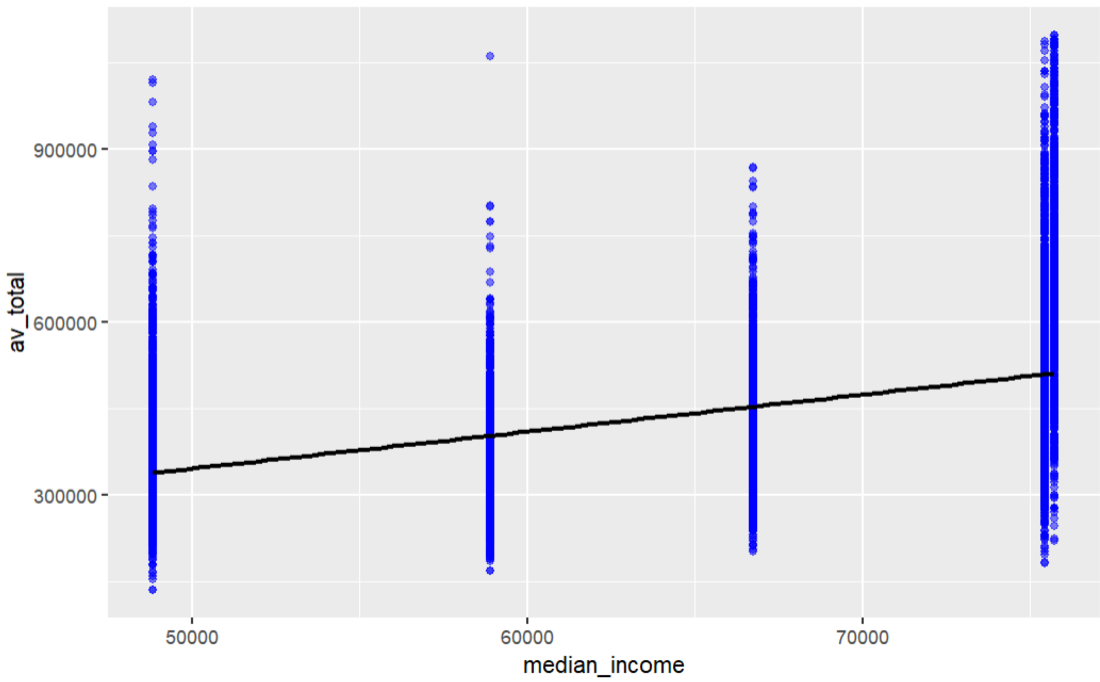

900000
600000
300000
av_total
0.0
2.5
5.0
7.5
10.0
r_half_bth


900000
600000
300000
av_total
1.0
1.5
2.0
2.5
3.0
r_kitch

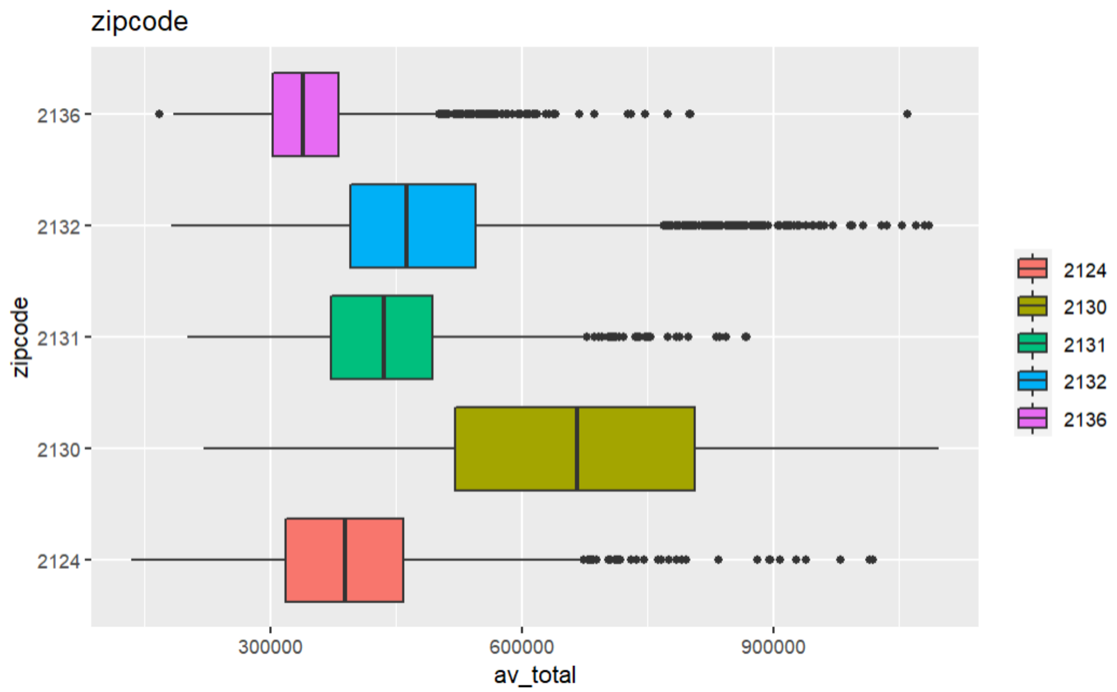

900000
600000
300000
av_total
0
1
2
3
4
5
r_fplace

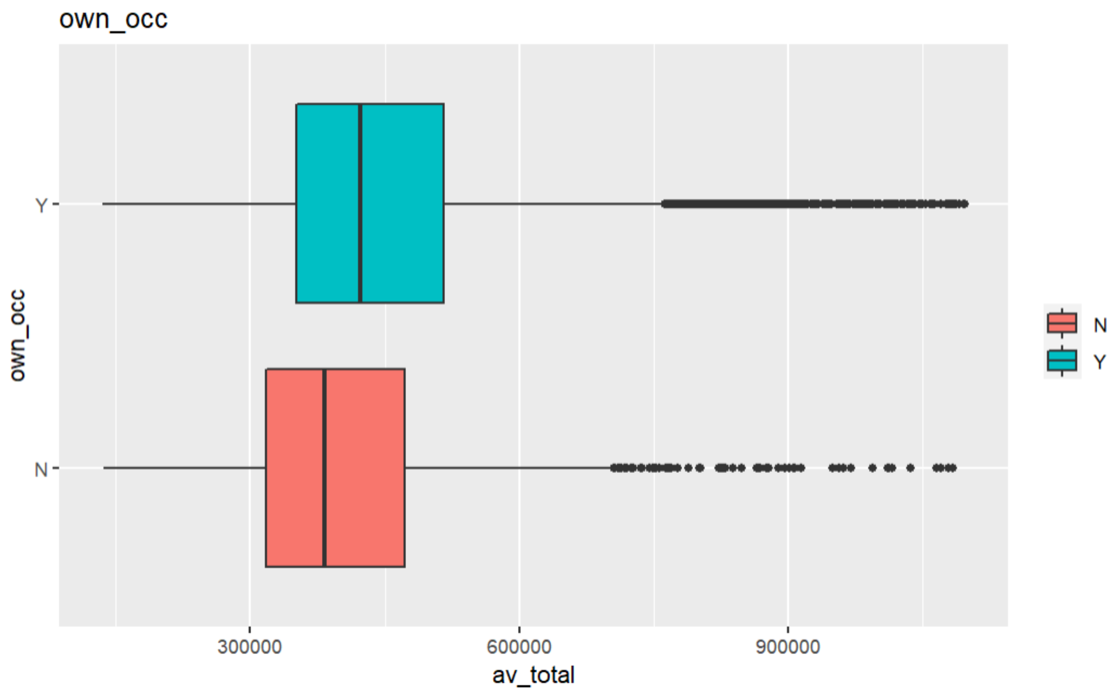

900000
600000
300000
av_total
30000
35000
40000
45000
population

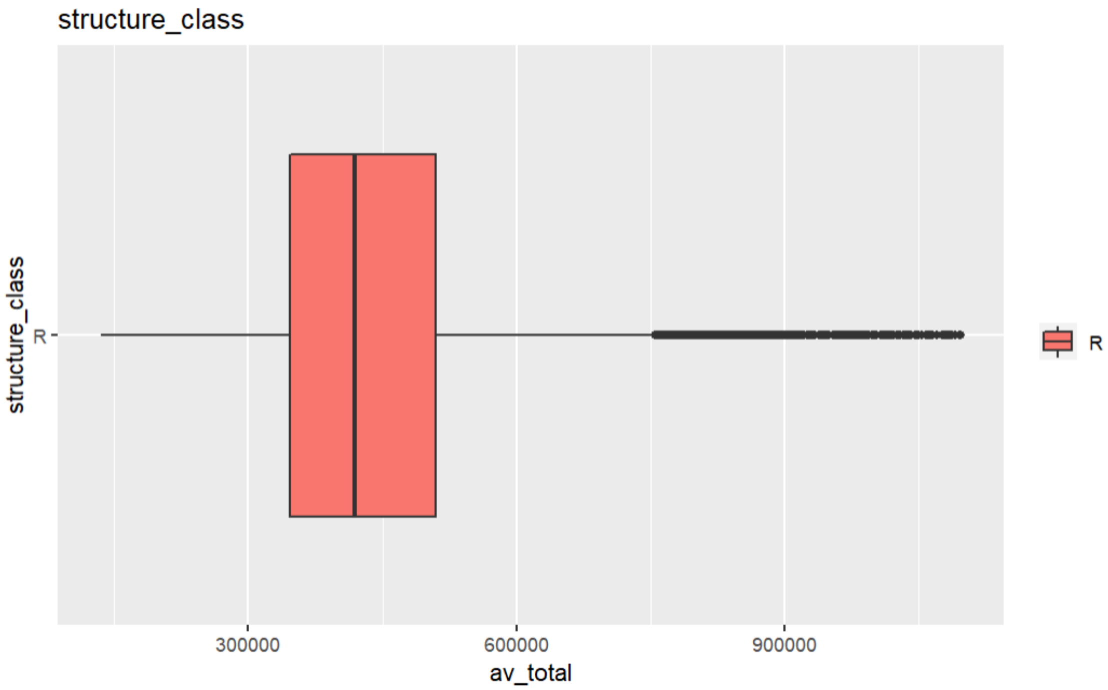


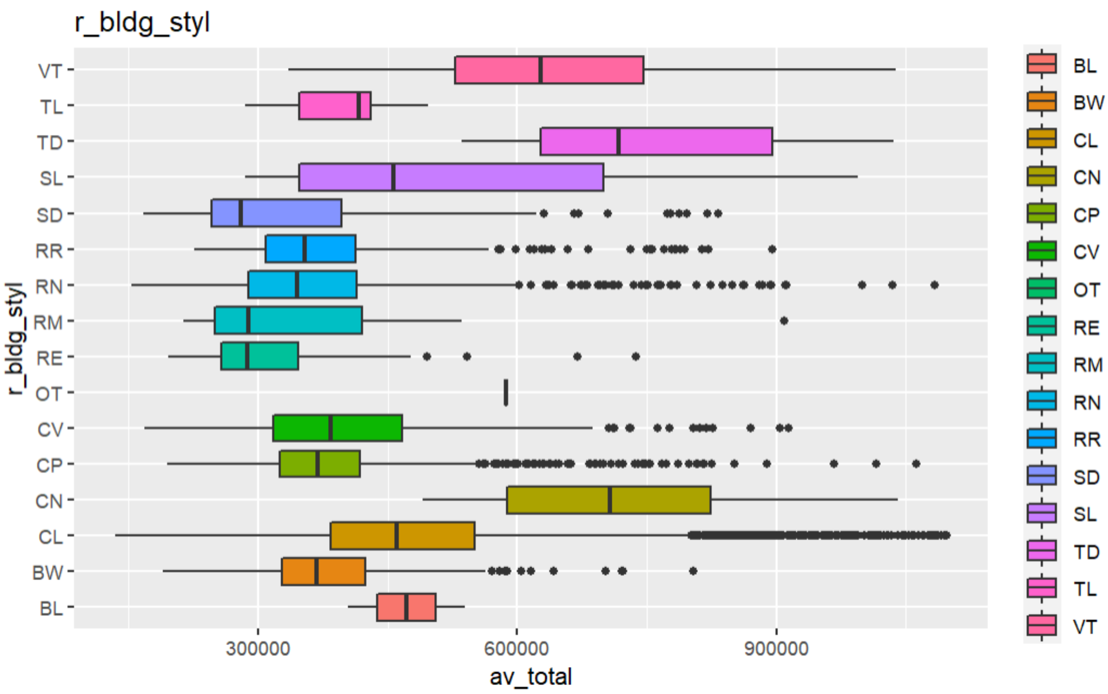


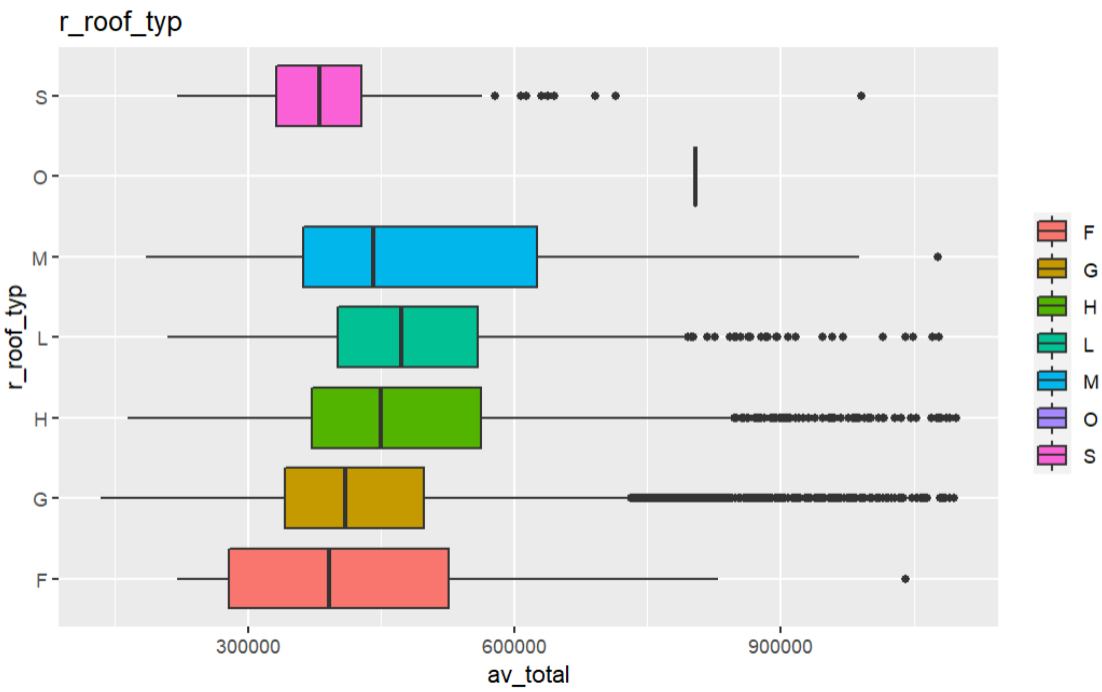
(2) Categorical Variables: zipcode, own_occ, structure_class, r_bldg_styl, r_roof_typ, r_ext_fin, r_bth_style, r_kitch_style, r_heat_typ, r_ac, r_ext_cnd, r_ovrall_cnd, r_int_cnd, r_int_fin, r_view, zip, city_state

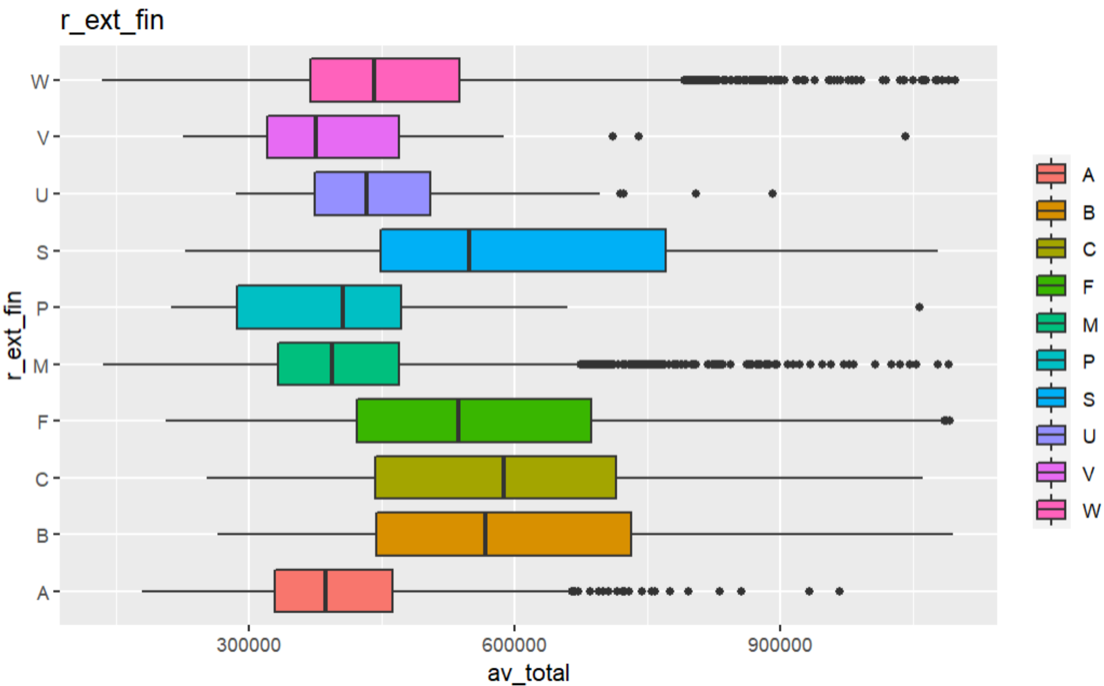

zipcode
2136
2132
2131
2130
2124
zipcode
300000
600000
900000
av_total
2124
2130
2131
2132
2136

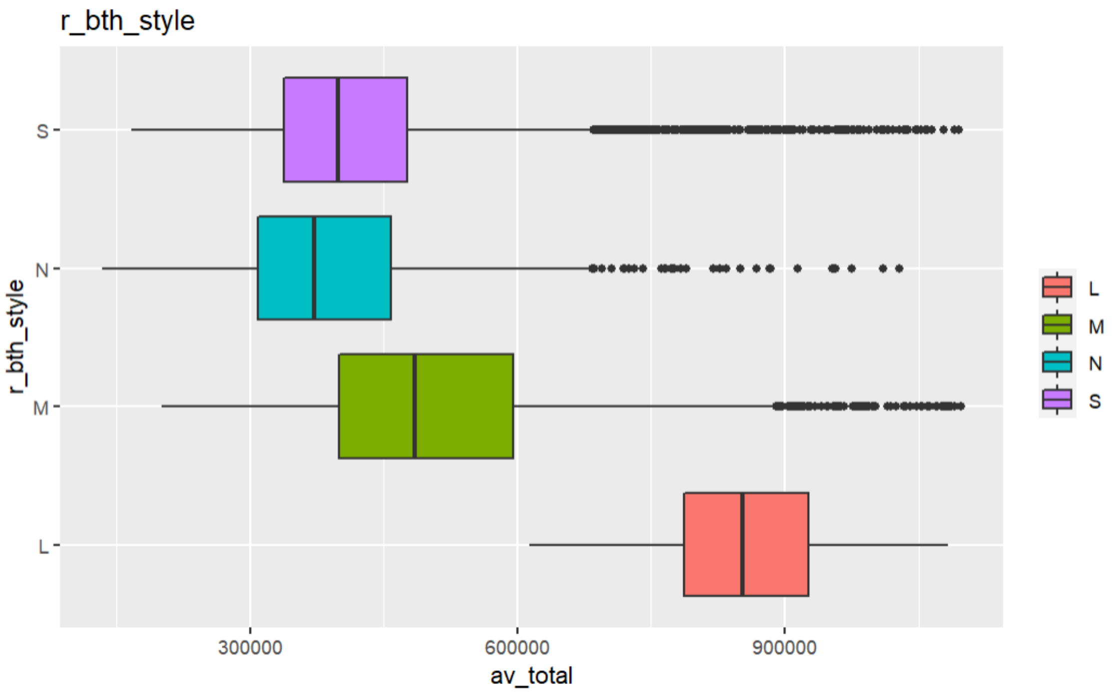

own_occ
Y
N
own_occ
300000
600000
900000
av_total
N
Y

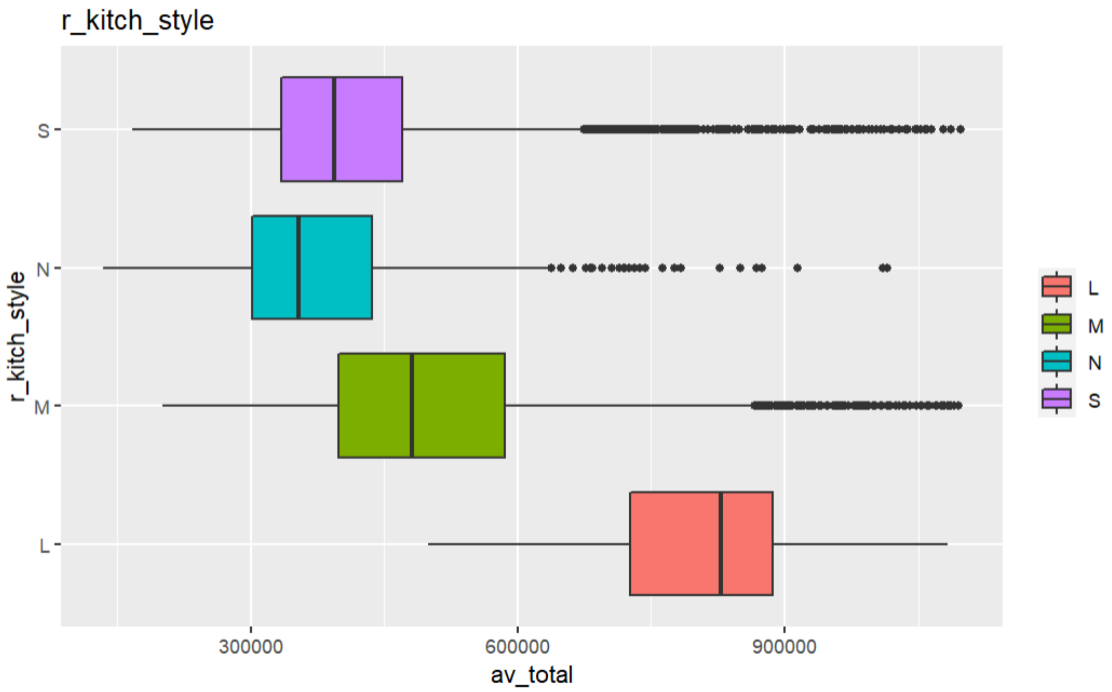

structure_class
structure_class
R
300000
600000
900000
av_total
R

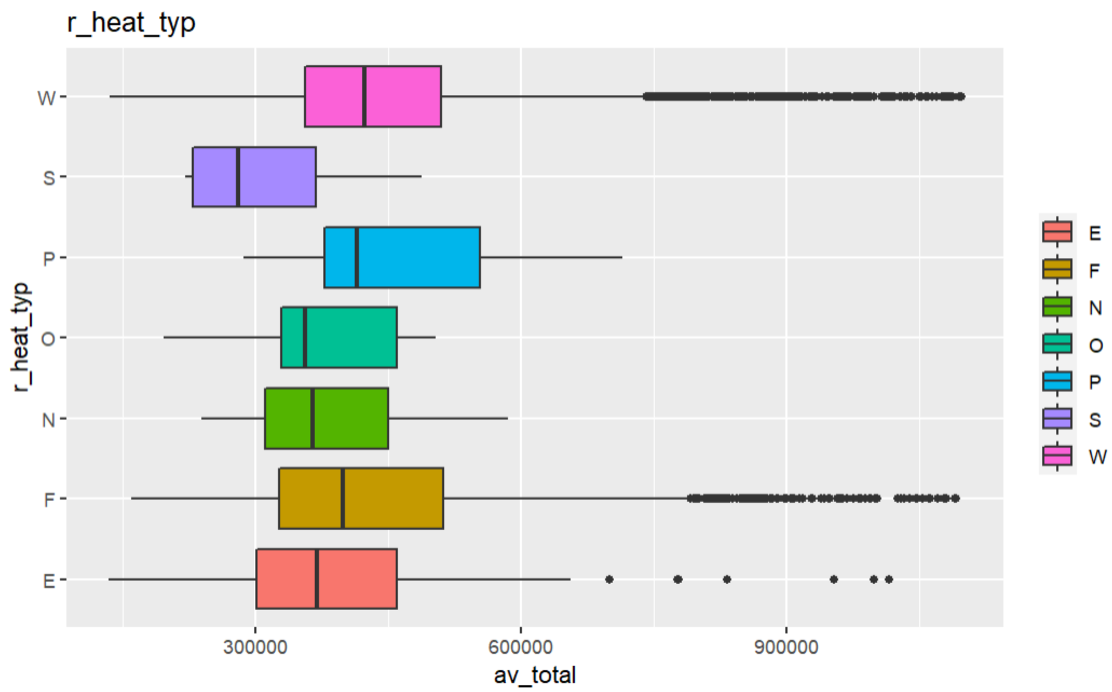

r_bldg_styl
r_bldg_styl
VT
TL
TD
SL
SD
RR
RN
RM
RE
OT
CV
CP
CN
CL
BW
BL
300000
600000
900000
av_total
BL
BW
CL
CN
CP
CV
OT
RE
RM
RN
RR
SD
SL
TD
TL
VT


r_roof_typ
S
O
M
L
H
G
F
r_roof_typ
300000
600000
900000
av_total
F
G
H
L
M
O
S

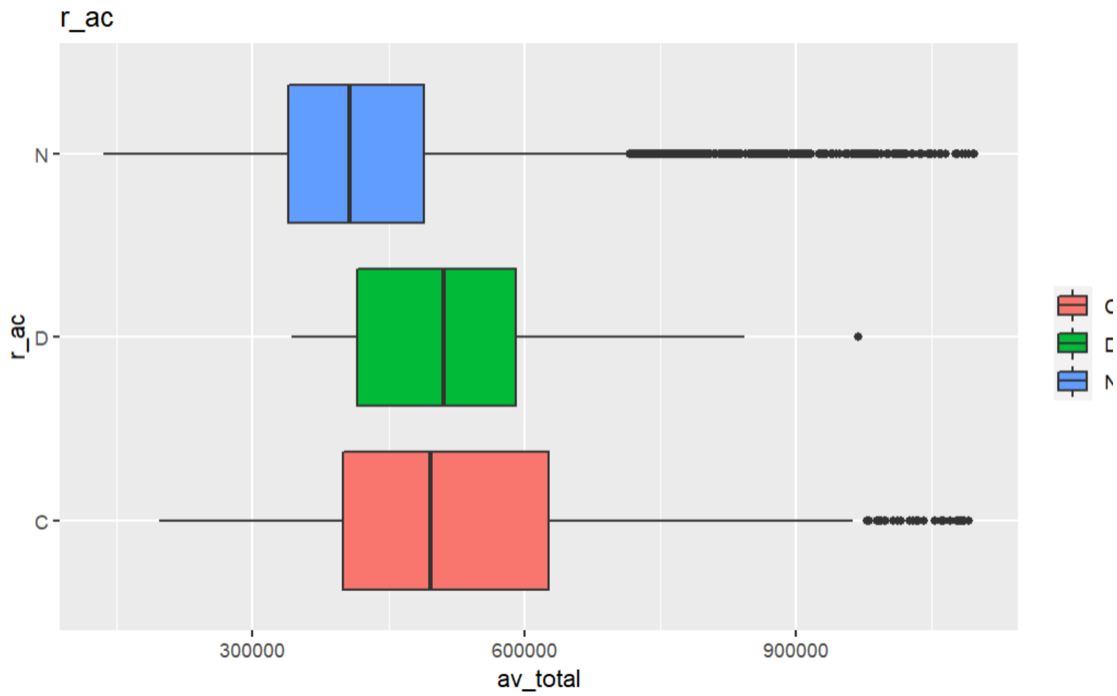

r_ext_fin
W
V
U
S
P
M
F
C
B
A
r_ext_fin
300000
600000
900000
av_total
A
B
C
F
M
P
S
U
V
W


r_bth_style
S
N
M
L
r_bth_style
300000
600000
900000
av_total
L
M
N
S

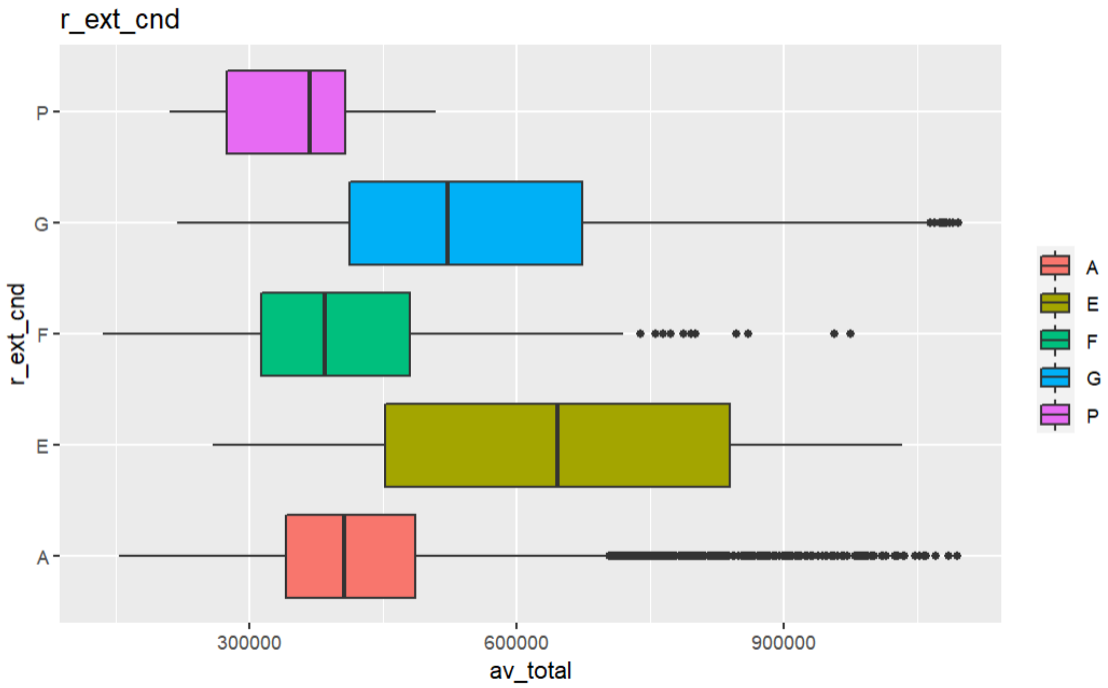

r_kitch_style
S
N
M
L
r_kitch_style
300000
600000
900000
av_total
L
M
N
S

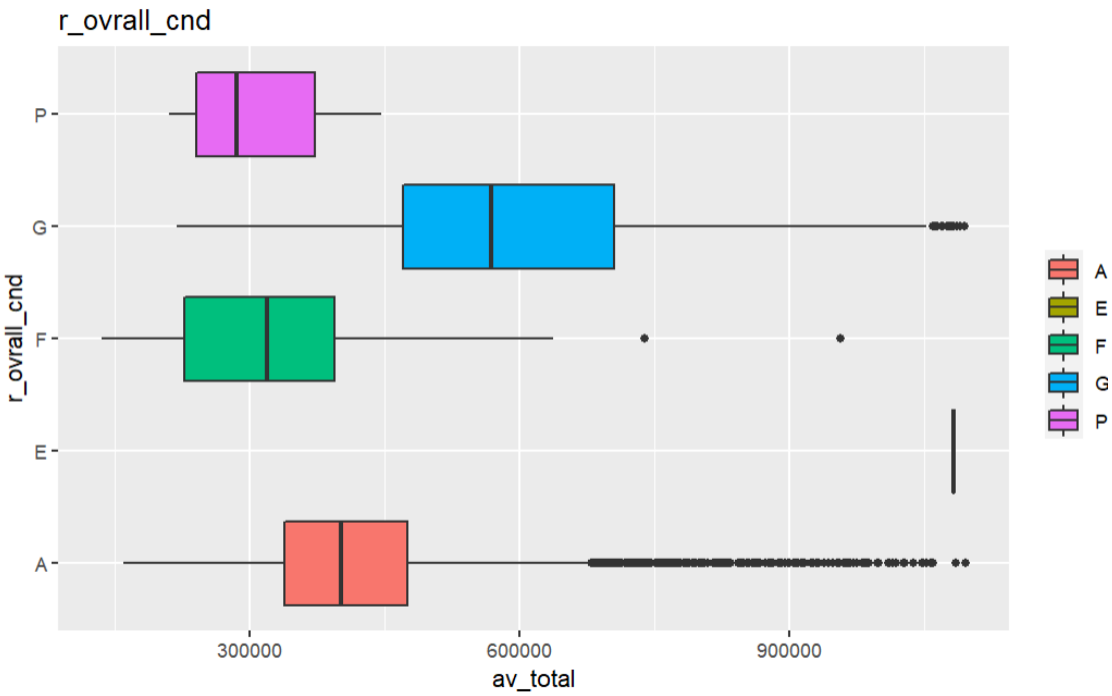


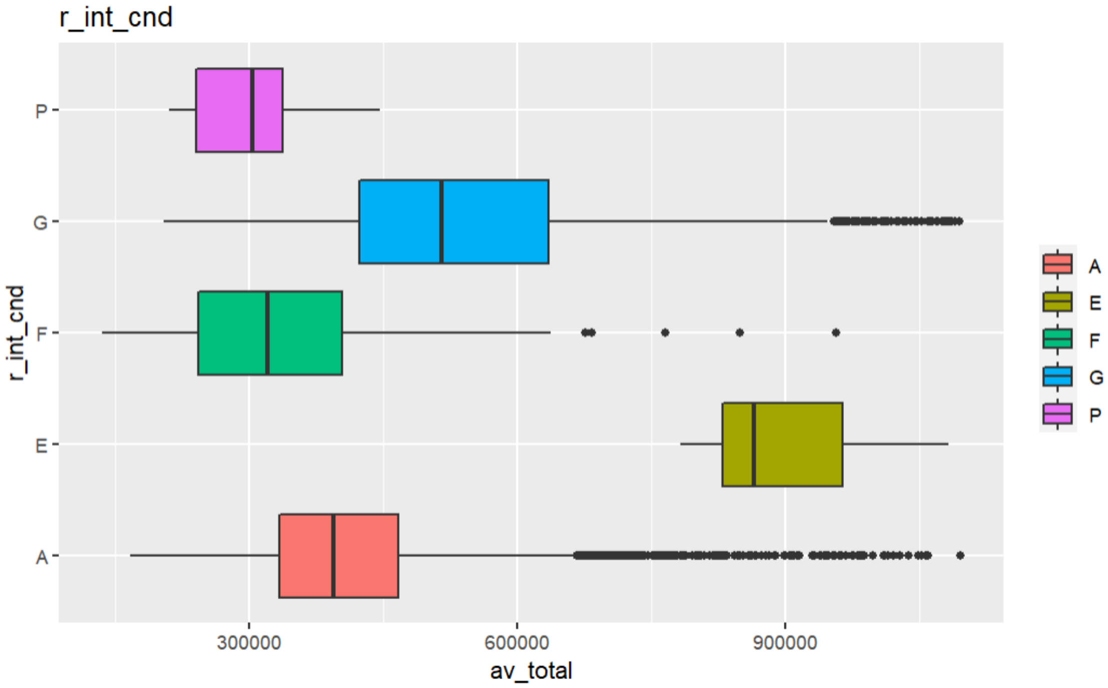
r_heat_typ
W
S
P
O
N
F
E
r_heat_typ
300000
600000
900000
av_total
E
F
N
O
P
S
W

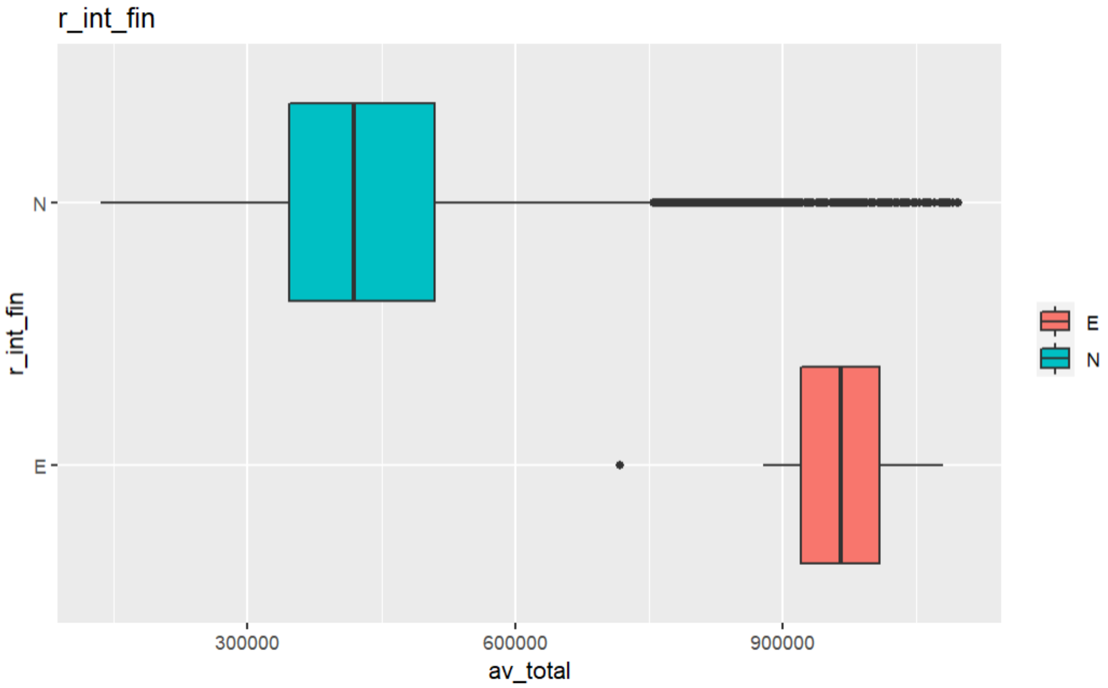


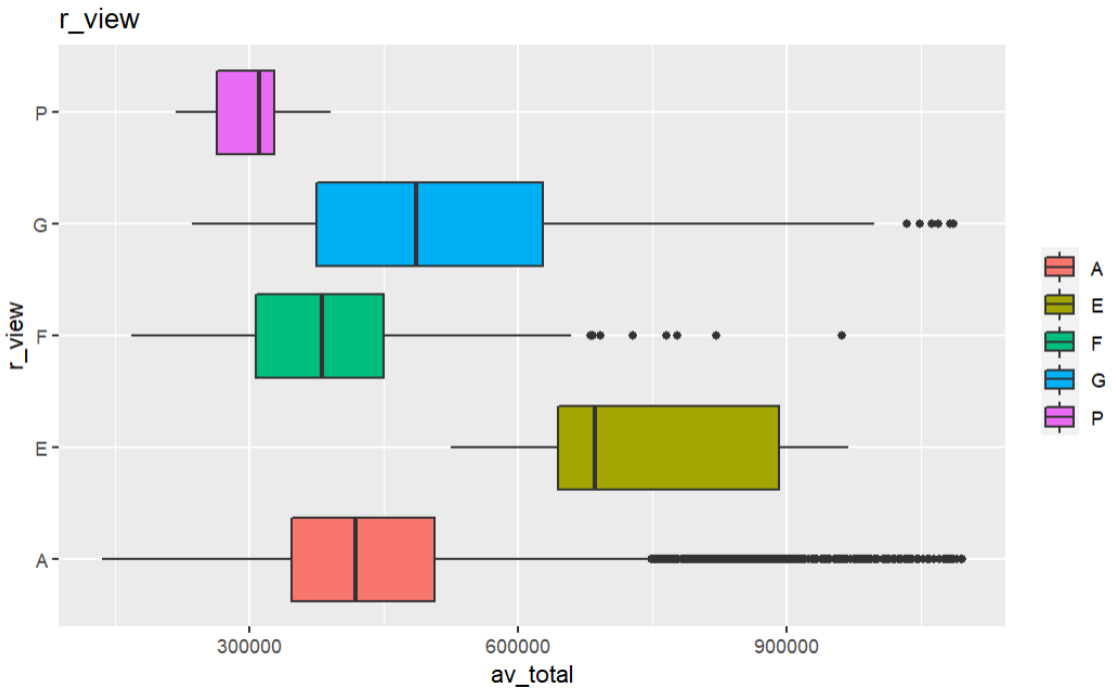
r_ac
N
D
C
r_ac
300000
600000
900000
av_total

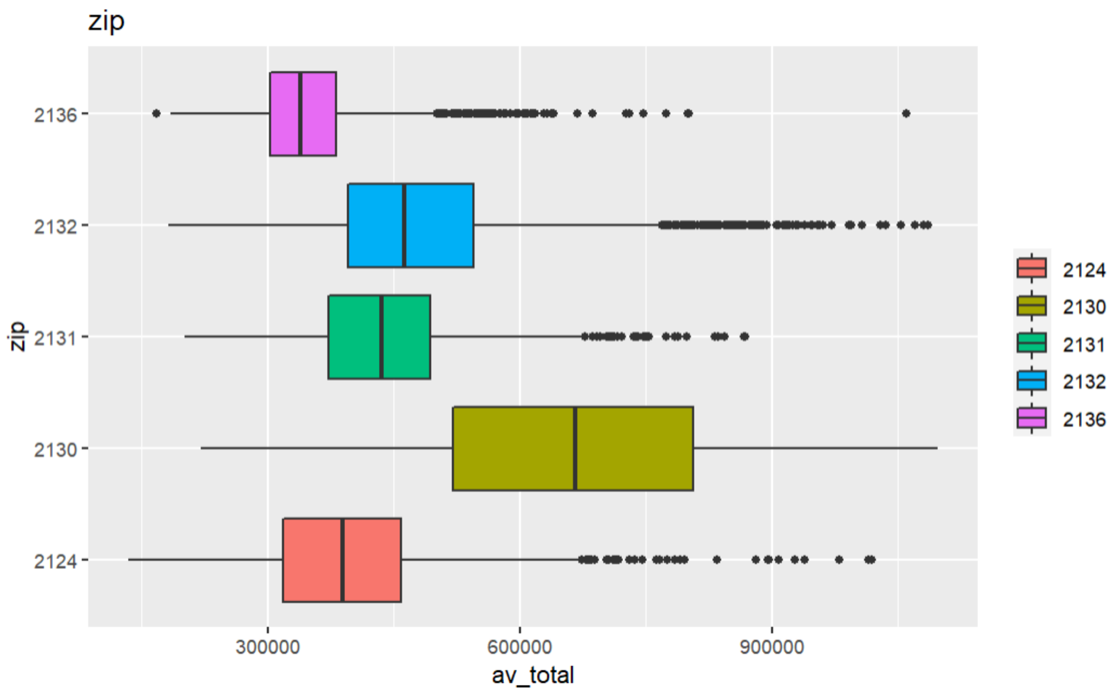

r_ext_cnd
P
G
F
E
A
r_ext_cnd
300000
600000
900000
av_total
A
E
F
G
P

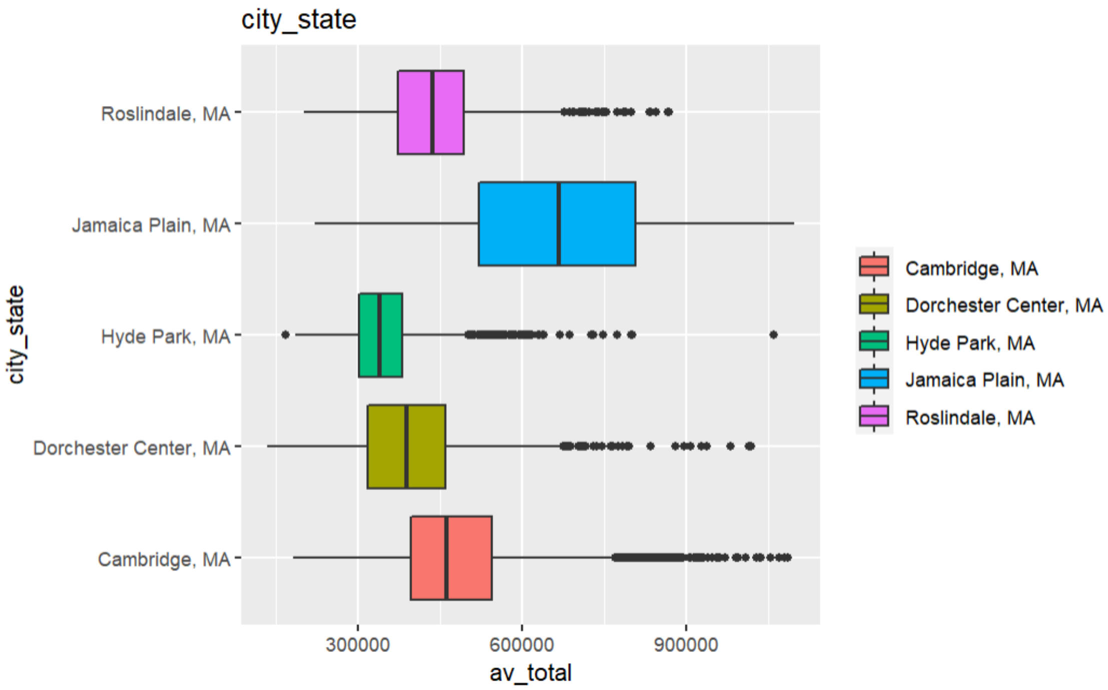

r_ovrall_cnd
P
G
F
E
A
r_ovrall_cnd
300000
600000
900000
av_total
A
E
F
G
P


r_int_cnd
P
G
F
E
A
r_int_cnd
300000
600000
900000
av_total
A
E
F
G
P


r_int_fin
N
E
r_int_fin
300000
600000
900000
av_total
E
N


r_view
P
G
F
E
A
r_view
300000
600000
900000
av_total
A
E
F
G
P


zip
2136
2132
2131
2130
2124
zip
300000
600000
900000
av_total
2124
2130
2131
2132
2136



9. Top & Bottom 10 predictions from your test set.

a. Top 10 predictions what is similar

No.	.pred	av_total	error	abs_error
1	597797.2	597800	2.8125	2.8125
2	380445.9	380400	-45.875	45.875
3	467848.1	467900	51.875	51.875
4	373419.7	373500	80.34375	80.34375
5	873613.5	873700	86.5	86.5
6	325804.6	325900	95.375	95.375
7	339899.5	340000	100.4688	100.4688
8	746579.4	746700	120.625	120.625
9	297775	297900	125	125
10	286957	287100	142.9688	142.9688

b. Bottom 10 predictions what is similar about these

No.	.pred	av_total	error	abs_error
1	800826.5	463200	-337626.5	337626.5
2	960118.8	641600	-318518.8	318518.8
3	909488.6	609800	-299688.6	299688.6
4	476887.7	767500	290612.3	290612.3
5	628724.4	363300	-265424.4	265424.4
6	838793	1090500	251707	251707
7	737873.9	981000	243126.1	243126.1
8	947911.6	705400	-242511.6	242511.6
9	943679.7	701400	-242279.7	242279.7

10	673414.6	443000	-230414.6	230414.6

10. Kaggle Submission

Kaggle Name: Eagle Xuhui Ying

Kaggle reported score: 51978.65534