1 Outlier Detection

1.1 Detect Outliers by Data Summary

We find an obvious outlier by checking summary of the data:

```
   RentTotal         SqftLease          Renovation       LeaseLen          Age           DistCity
 Min.   : -50184   Min.   :   211.6   Min.   : 0.00   Min.   : 3.00   Min.   : 0.00   Min.   :0.100
 1st Qu.:  99434   1st Qu.:  5308.6   1st Qu.: 7.00   1st Qu.: 4.00   1st Qu.:13.00   1st Qu.:0.950
 Median : 226898   Median : 11379.2   Median :15.00   Median : 6.00   Median :29.00   Median :2.110
 Mean   : 469459   Mean   : 22739.6   Mean   :19.84   Mean   : 6.08   Mean   :30.52   Mean   :2.492
 3rd Qu.: 595084   3rd Qu.: 30373.9   3rd Qu.:32.00   3rd Qu.: 8.00   3rd Qu.:44.00   3rd Qu.:3.930
 Max.   :3673653   Max.   :198043.1   Max.   :76.00   Max.   :10.00   Max.   :83.00   Max.   :6.370
    DistAirp        DriveAirp       Location    Occupancy        FloorsBldg      SqftFloor
 Min.   :10.07   Min.   :12.32   CITY  :90   Min.   :0.3100   Min.   : 3.00   Min.   : 6146
 1st Qu.:12.46   1st Qu.:21.02   SUBNEW:82   1st Qu.:0.7900   1st Qu.:11.00   1st Qu.:23766
 Median :13.94   Median :24.76   SUBOLD:53   Median :0.8900   Median :16.00   Median :32297
 Mean   :14.21   Mean   :26.01               Mean   :0.8501   Mean   :18.19   Mean   :32649
 3rd Qu.:15.37   3rd Qu.:29.68               3rd Qu.:0.9500   3rd Qu.:21.00   3rd Qu.:40684
 Max.   :20.45   Max.   :49.37               Max.   :1.0000   Max.   :74.00   Max.   :70445
   Elevators      Restaurant Wiring    Exercise     DistHosp        FirmType    FloorLease
 Min.   : 2.000   NO :158    NO :188   NO :194   Min.   :0.0400   BUS   :70   Min.   : 1.000
 1st Qu.: 4.000   YES: 67    YES: 37   YES: 31   1st Qu.:0.5600   DOCTOR:48   1st Qu.: 3.000
 Median : 5.000                                 Median :0.8900   GOVT  :25   Median : 8.000
 Mean   : 5.582                                 Mean   :0.9996   LEGAL :57   Mean   : 9.213
 3rd Qu.: 7.000                                 3rd Qu.:1.4300   OTHER :25   3rd Qu.:13.000
 Max.   :13.000                                 Max.   :2.5800               Max.   :47.000
 Renewable    Parking
 NO :194   Min.   : 0.000
 YES: 31   1st Qu.: 0.000
           Median : 2.000
           Mean   : 3.244
           3rd Qu.: 4.000
           Max.   :23.000
```

RentTotal should not be negative number, so delete the observation with a RentTotal less than 0.

1.2 Detect Outliers by Mahalanobis Distance

Because we have multiple predictors here, so when we decide if an observation is an outlier, we'd better to collectively consider multiple variables that matter.

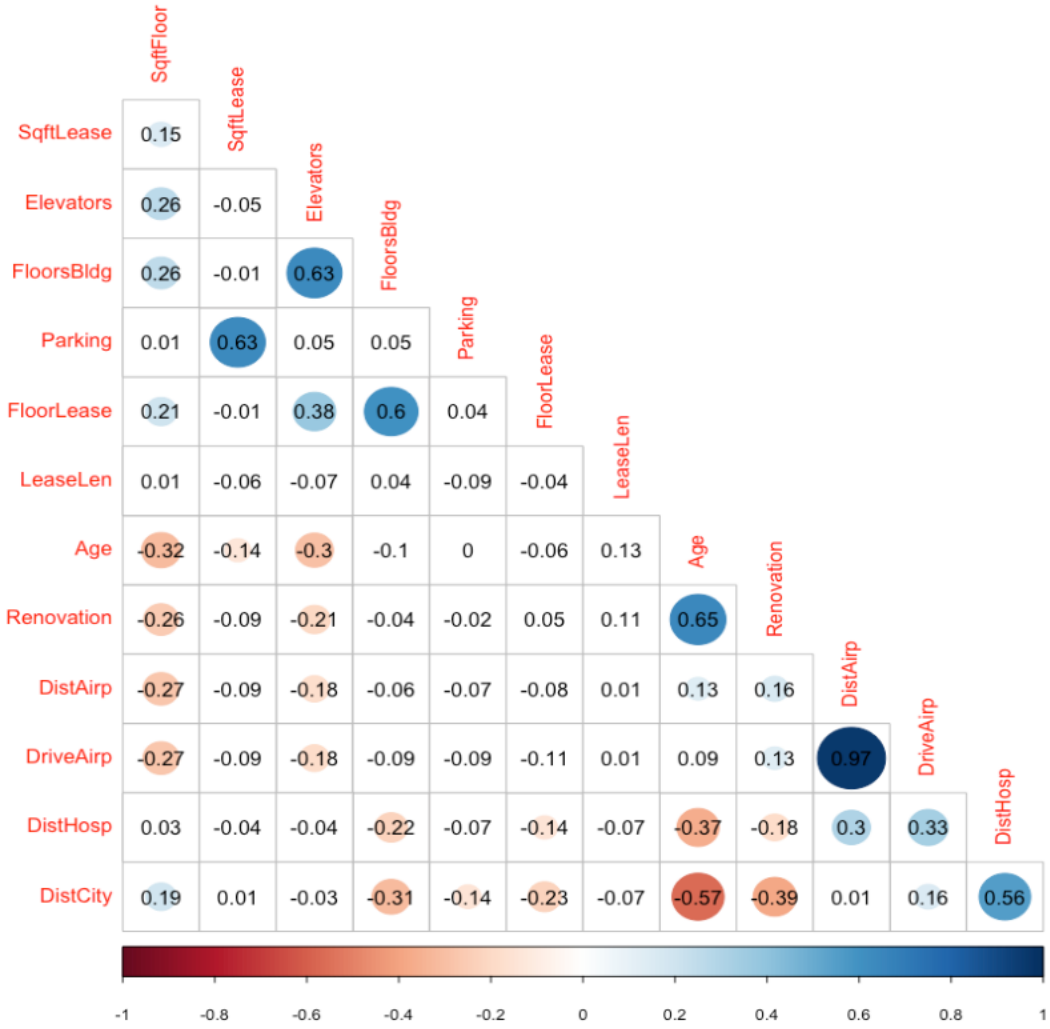
So, I choose to detect outliers based on Mahalanobis distance, which allows us to declare an observation as an outlier based on all continuous predictors. (Basically, Mahalanobis distance is a multi-dimensional generalization of outlier detection by Z-score.)

By implementing this method, I detect 4 outliers: observations 20, 21, 217 and 219, and delete them from the data.

	SqftLease <dbl>	Renovation <int>	LeaseLen <int>	Age <int>	DistCity <dbl>	DistAirp <dbl>	FloorsBldg <int>	SqftFloor <dbl>	Elevators <int>	DistHosp <dbl>	FloorLease <int>	Parking <int>
20	23747.04	1	6	1	0.10	14.50	73	47424.05	12	0.79	9	1
21	198043.06	14	8	14	1.94	15.93	12	38062.87	3	1.55	7	23
217	153277.04	8	10	8	5.52	10.43	19	60560.46	3	0.16	8	0
219	6121.72	6	9	6	1.42	15.84	74	58272.26	10	0.48	47	2

2 Feature Engineering

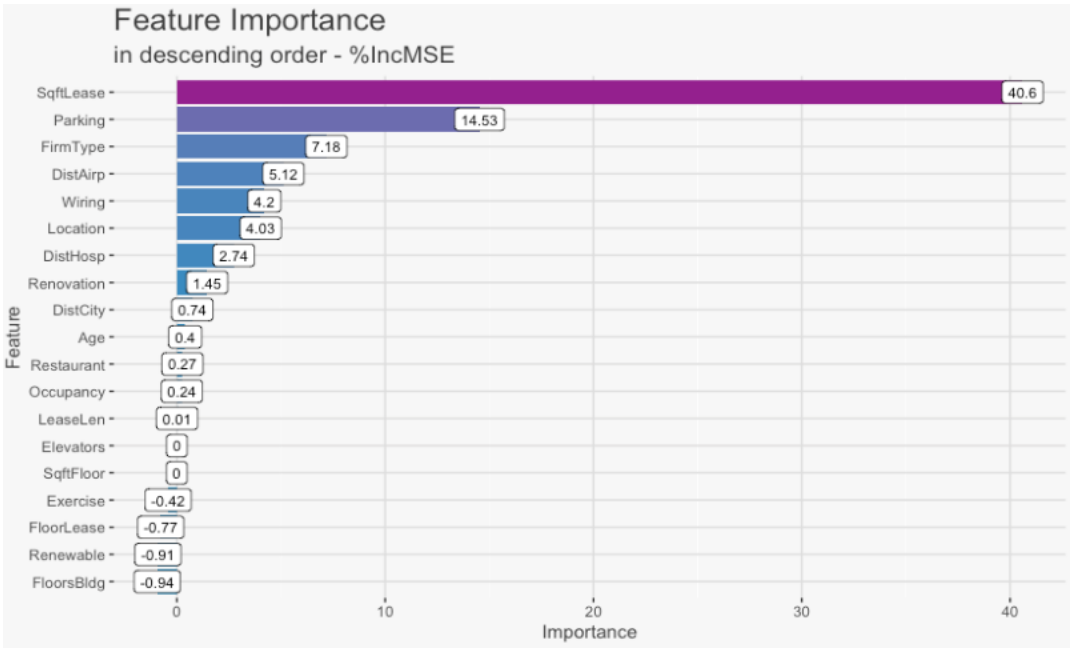
Highly corelated features might diffuse feature importance in later analysis, so firstly we need to inspect the correlation between numeric variables:



DistAirp and DriveAirp are highly correlated (correlation coefficient = 0.97). Thus, we delete variable DriveAirp.

3 Feature Importance (Q1)

I initial a random forest model to fit the data, and then compute the feature importance based on %IncMSE. It is the increase in MSE (Mean Squared Error) of predictions as a result of one variable being permuted. So, the higher %IncMSE of one variable is, the more important this variable is.

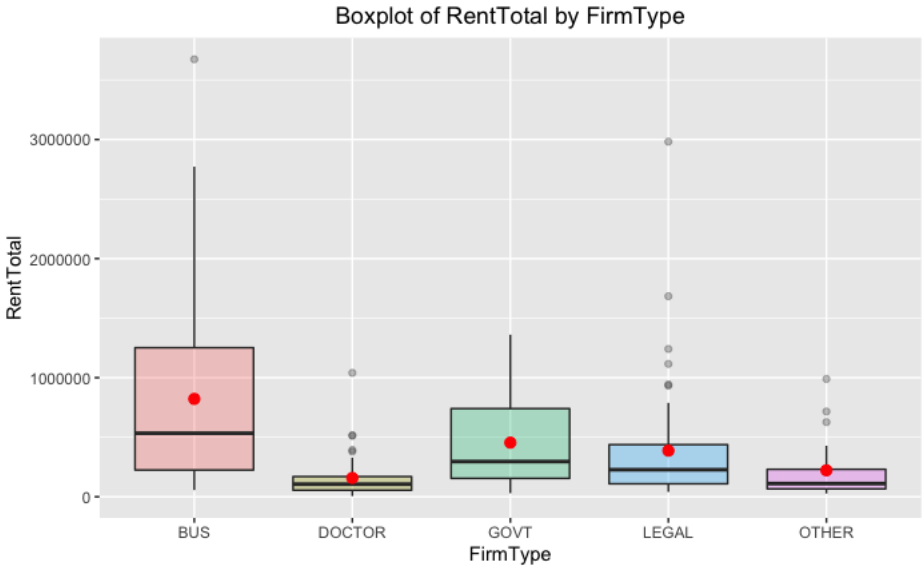


The top 3 factors that determine rent are SqftLease, Parking, and FirmType.

4 Impact of FirmType on RentTotal (Q2)

FirmType is an important feature(#3 in Feature Importance). It means choosing properties in the building with different majority type of firms has a great impact on rent.

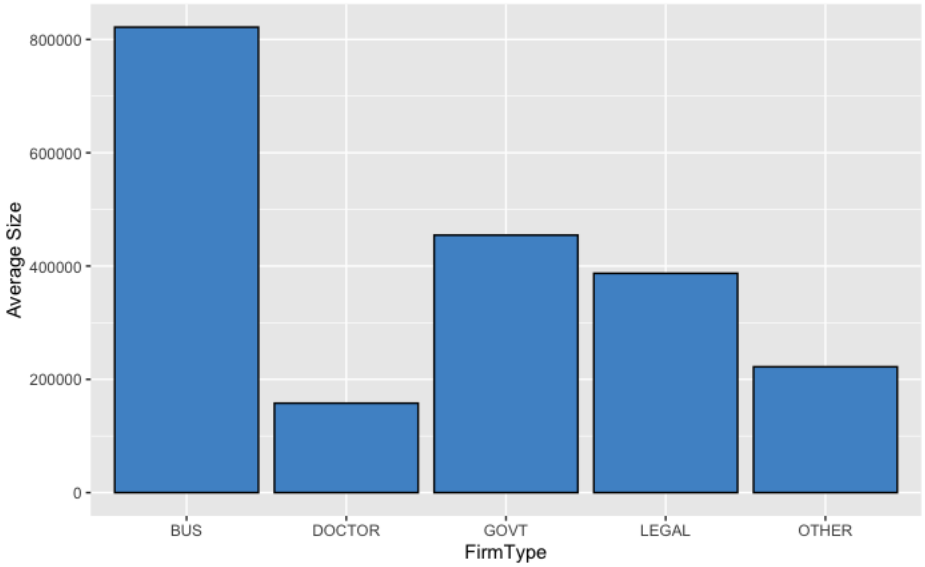
4.1 RentTotal by FirmType



(red point in the figure represents average rent)

With the highest average rent and the highest medium rent, properties in the building with majority Bus Firm tend to have higher rent, followed by Government Firm, Legal Firm, Other Firm, and Doctor Firm.

It's quite strange that properties in the building with majority Doctor Firm has the lowest rent while the one with Bus Firm has the highest. One possible explanation could be rent is also related to size of the lease.



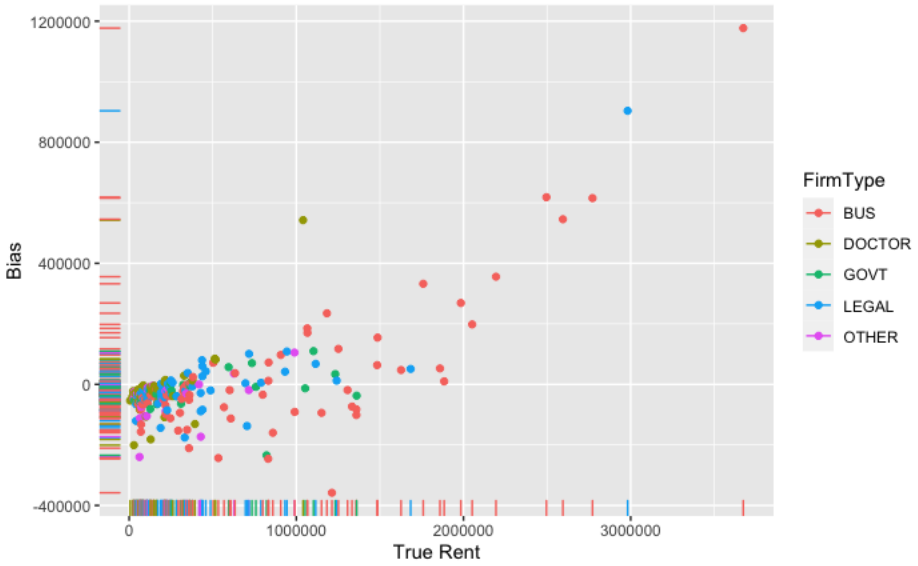
It's obvious that the properties in the building with a great number of Doctor Firm has much smaller average size of the lease than the one with majority Bus Firm, which verified our assumption above.

4.2 Residual Analysis

Predict on all data by the random forest model and calculate the bias:

Bias = Actual Rent – Predicted Rent

If the bias > 0, it means rent of this property is higher than it should be, and it is overpaid. Otherwise, bias < 0, means that property is underpaid.

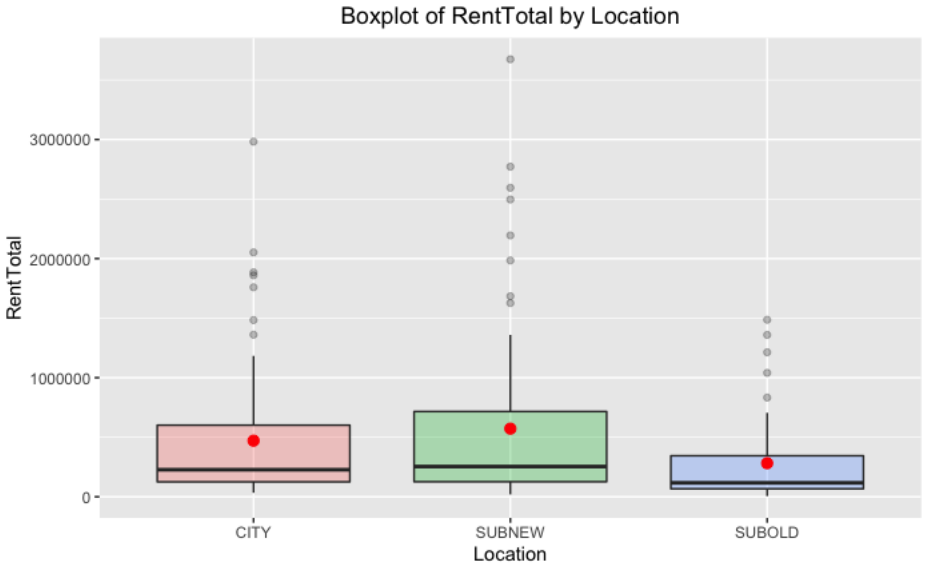


From the above figure, we can see that when majority type of firms in the building is Bus Firm, the property with high rent there are more likely to be overpaid.

5 Impact of Location on RentTotal (Q3)

Compared with the variable FirmType, Location is less important(#6 in Feature Importance), which means different locations influence less on rent of property.

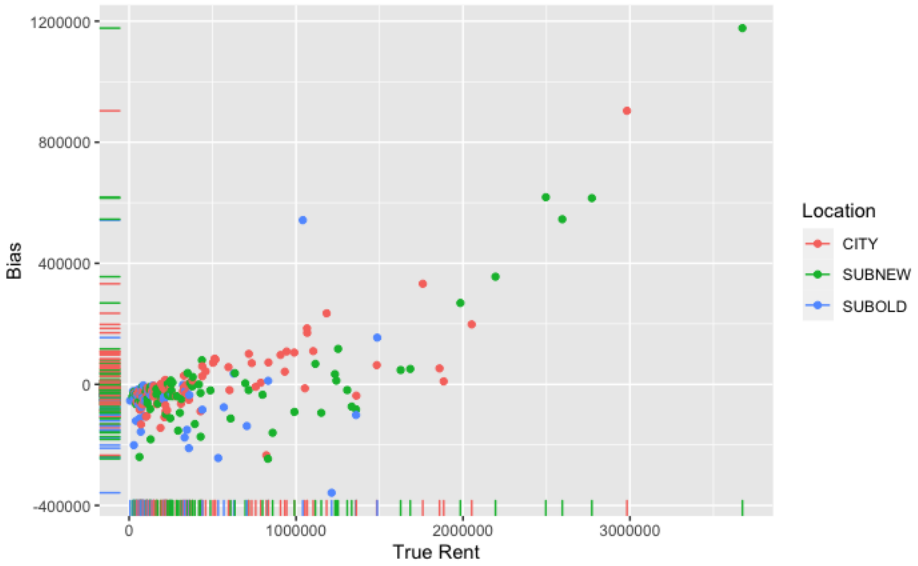
5.1 RentTotal by Location



(red point in the figure represents average rent)

Properties in new suburb tend to have higher rent, while properties in old suburb tend have lower rent.

5.2 Residual Analysis



The properties with high rent in city and new suburb tend to be overpaid.