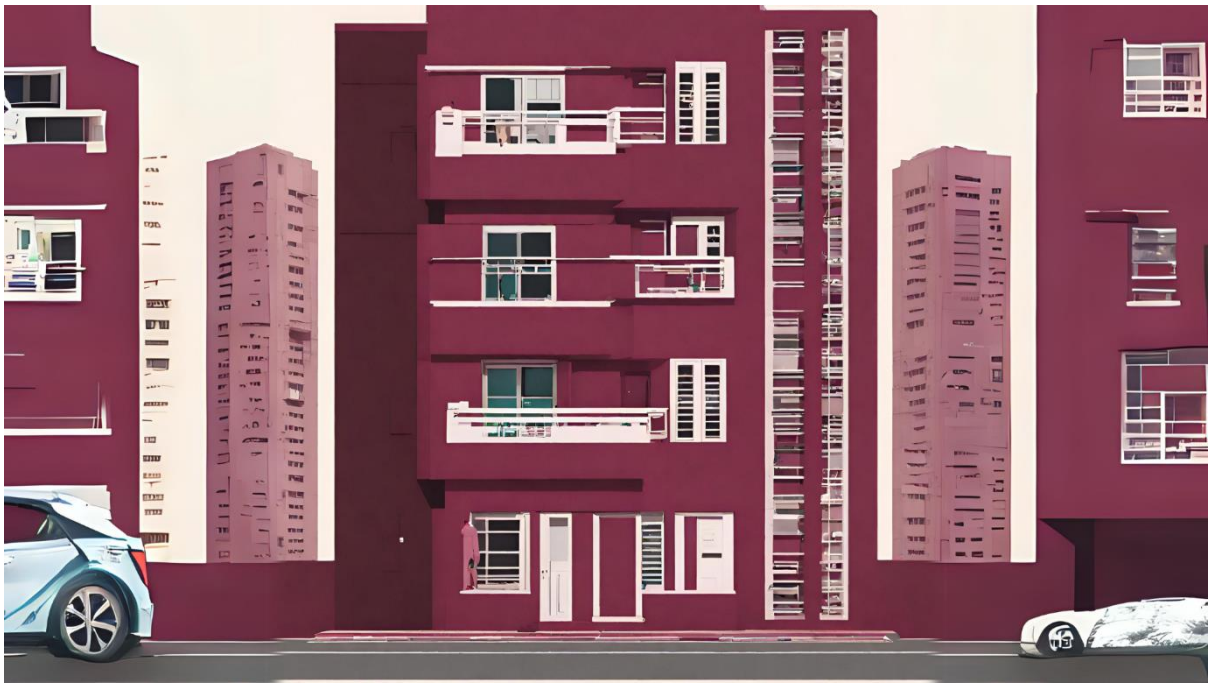


Singapore's HDB Resale Market

Programming and Data Analysis for Business
(LA E23 BINTV2006U)

Student number: 165243
20,978 Characters incl. spaces (9.22 pages)

Abstract

As a country facing land scarcity with a paternalistic government, Singapore's housing market is unique. As of 2022, 78.3% of Singaporeans live in Housing and Development Board (HDB) flats[1], which is Singapore's public housing. Buying and renting a typical HDB flat comes with a multitude of regulations, making the public housing market an interesting case to study.

Although public housing is intended to be affordable for the masses, Singapore's public housing prices has seen an uptick of million-dollar sales recently[2]. Henceforth, this paper aims to investigate *the factors that lead to a higher public housing resale price, and find a satisfactory model to predict the resale prices.*

Exploratory data analysis is used to briefly explore the relationships between variables and resale prices. Two supervised machine learning models, Linear Regression and Random Forest Tree, were used to model the relationships. RMSE is used to compare the performance of both models.

The results of the analysis should give insights into which factors boosts the resale price of HDB flats. It is recommended that consumers pay attention to factors such as location, size and remaining lease left for the flat. The Linear Regression model performed better with a lower RMSE than the Random Forest Tree model, with a RMSE of 42,346 compared to 25,547, and thus is able to give a better gauge of a HDB flat's resale price based on its attributes.

Introduction

Motivation

Singapore's public housing authority, Housing and Development Board (HDB), aims to provide Singaporeans with affordable and quality homes[3]. To keep HDB flats accessible for all, ownership of these flats is limited to a 99-year leasehold. Under strict regulations pertaining reselling and ownership, HDB flats are comparatively priced lower than private houses.

Many Singaporeans own an HDB flat, creating an unspoken expectation amongst many voters to ensure HDB flats appreciate. There is tension between ensuring prices stay within the reach of Millennials and Gen Zs, while allowing Gen Xs and Baby Boomers to profit from the sale of their HDB flats. To balance opposing interests, the government has intertwined multiple policies. Through the investigation of relationships between variables and resale prices, the paper discusses governmental policies that have influenced resale prices. Ultimately, an understanding will be reached on factors that help fetch a higher HDB flat resale price, and find a model to reasonably predict resale prices.

Methods

Data Inspection

The data on HDB resale prices is taken from Singapore's national open data website, and is split by years. The date range of the data is from March 2012 to November 2023, which is over 11 years of data. In total, I have 255,405 observations across three sets of data, each of separate year ranges. However, I will use sample_n to take only 25,000 observations.

---

[1] https://www.singstat.gov.sg/find-data/search-by-theme/households/households/latest-data
[2] https://www.reuters.com/markets/asia/singapore-sees-rise-million-dollar-public-housing-2022-08-31/
[3] https://www.hdb.gov.sg/cs/infoweb/about-us

One of the year ranges has only 10 variables, while the other two has 11. The data from March 2012 to December 2014 does not have a remaining_lease column. The data from January 2017 5o November 2023 has a remaining_lease column in terms of years and months, but that of January 2015 to December 2016 is in terms of years only. Hence some standardization is needed here, and I chose to keep the remaining_lease column but in terms of years only, as it is difficult to find in terms of months since we only know the year of lease commencement. The following code is a sample of the standardization:

```
Mar2012toDec2014 <- Mar2012toDec2014 %>%
        mutate(sale_year = as.numeric(str_extract(Mar2012toDec2014$month, "^.{4}"))) %>%
        mutate(remaining_lease = 99 - (sale_year - lease_commence_date)) %>%
        select(-sale_year)
```

I then used rbind to combine the different years data into one data frame for wrangling.

Feature Engineering

Singapore's education topped the world in 2022, according to Pisa[4]. At seven, children are enrolled into primary school for six years, before moving on to secondary and tertiary education. Admission into secondary schools and beyond is through one's score at national exams pitted against everyone else's. But for primary schools, admission is through balloting if there are insufficient slots for the number of applications. There are extra chances awarded under certain concessions, such as residing within 2km of the school[5]. With this, a prevalent mindset exists within parents to enrol their child into a good primary school for a head start. As such, I want to investigate if the number of primary schools within 2km of a HDB flat will influence its resale price.

Another consequence of Singapore's land scarcity is insufficient space on roads. Amidst other reasons like environmental consciousness, Singapore also aims to go car-lite. To do so, the cost of a new Toyota Corolla Sedan is almost three times as much in Singapore than Copenhagen[6]. Conversely, the average price of a public transport ticket in Singapore is about half that of Copenhagen's. Thus, I theorise that the availability of Mass Rapid Transit (MRT) stations (which are metro stations) within walking distance of a house will increase its resale value. Using an average walking speed for a walking duration of 15 minutes, within walking distance is 1km.

Hence, I pulled in data others have scrapped to add two variables – number of primary schools within 2km, and number of MRT stations within 1km (see Appendix B).

Standardization of Extra Data

To add the variables, I first found the latitude and longitude of each observation of the HDB data. I found latitude and longitude data of every address in Singapore and used left join to combine this data set with the HDB data.

Before combining, I noticed that there are differences in the formatting of the address column between the HDB data and the Zip Code data. The HDB data shortens various words like 'BLOCK' to 'BLK' and 'ROAD' to 'RD'. Hence, I shortened the address column in the Zip Code data since the HDB data is much larger in size (see Appendix A).

[4] https://www.businesstimes.com.sg/international/singapore-leads-way-asia-tops-world-education-class
[5] https://www.moe.gov.sg/primary/p1-registration/understand-balloting
[6] https://www.numbeo.com/cost-of-living/compare_cities.jsp?country1=Denmark&city1=Copenhagen&country2=Singapore&city2=Singapore

After combining, each observation of the HDB data will have its latitude and longitude coordinates. Next, using the Haversine formula, I created a max_lat function to find the maximum latitude given a distance from a point. To match primary schools under 2km of distance to each HDB listing in the Working Data, I first sorted both data frames by latitude.

I then used a loop to match all primary schools under the 2km distance from each HDB listing. I used the same process to add the MRT stations within 1km.

```
for (i in 1:nrow(WorkingData)) {
  max_lat = max_lat_fun(WorkingData[i,13])

  for (x in 1:nrow(PrimarySchSg)) {

    if (PrimarySchSg[x,7] < max_lat) {

      distance_btwn = distHaversine(c(WorkingData[i, 14], WorkingData[i,13]),
       c(PrimarySchSg[x,8], PrimarySchSg[x,7]))

      if (distance_btwn < 2000) {

        if (is.na(WorkingData[i, 16])) {
          WorkingData[i,16] = PrimarySchSg[x, 1]
        }

        else if (WorkingData[i,16] != PrimarySchSg[x, 1]) {
          WorkingData[i,16] = paste0(WorkingData[i, 16], ', ', PrimarySchSg[x, 1])
        }
      }
    }
  }
}
```

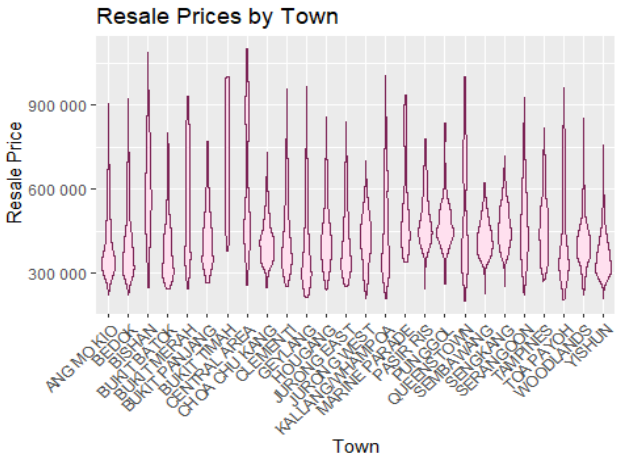
Some cells had more than one value. Hence, I used str_split to split the Primary School and MRT Station columns, before using mutate to add another column that sums up the number of facilities nearby each HDB listing.

Exploratory Data Analysis

Checking the normality of the resale price, the variable we want to predict, they are right-skewed, with a greater concentration of houses being sold at the lower end of the range. Most houses remain affordable, but there is a small yet significant number sold at high prices to skew the distribution. As of August 2023, 2.2% of resale flat transactions were million-dollar flats[7].

House Prices by Town

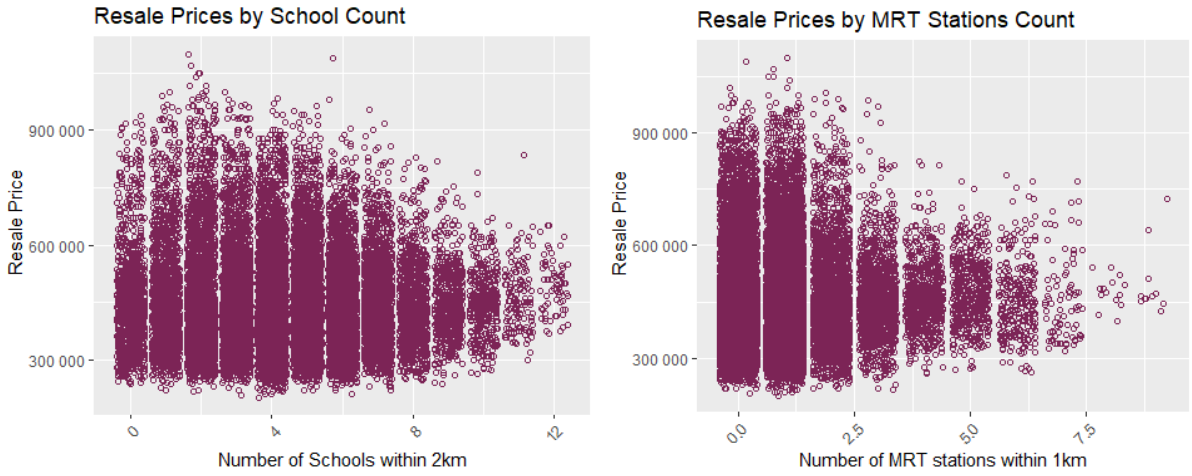
Towns such as Bishan, Central Area, Kallang and Queenstown have higher upper limits. These areas are located near the south of Singapore), where the Central Business District (CBD) resides. These towns are considered 'prime' locations and offer great connectivity to the rest of the island. Residing in these towns offer convenience and time-saving in commuting between home and work, thus houses in such areas command a premium.



[7] https://stackedhomes.com/editorial/why-are-there-so-many-millon-dollar-hdb-flats-still/#gs.1rvsjr

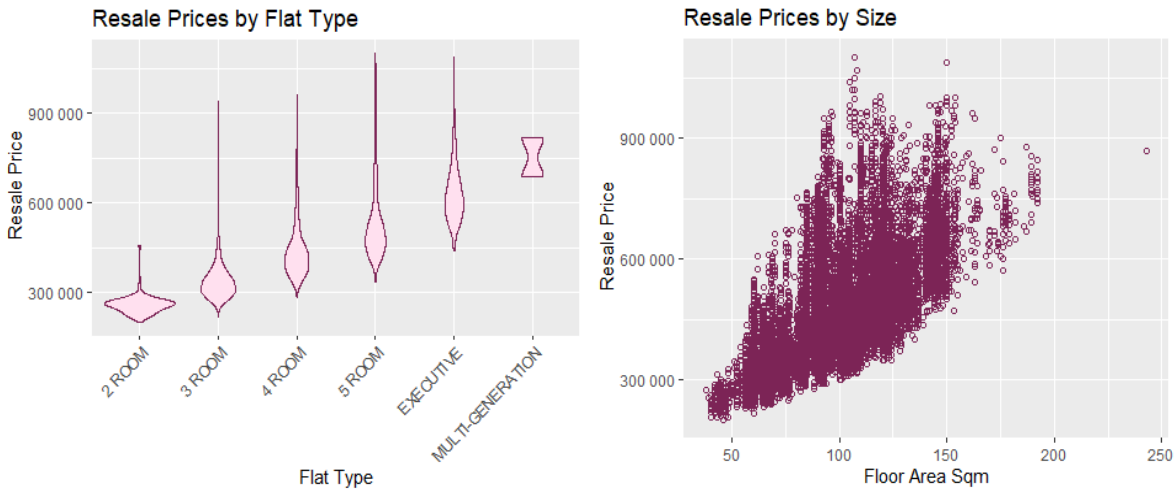
Another aspect is the maturity of these towns, an attribute describing how developed the town is. Mature towns have more amenities such as transport networks, shopping malls, schools, and parks[8]. These towns are typically older and hence well established, as Queenstown is an example of one of the oldest residential towns[9].

House Prices by Amenities (Primary Schools and MRT Stations)



Majority of HDB houses have at least 1 school within 2km and an MRT station within 1km. However, it seems that resale price does not necessarily increase with the number of primary schools and MRT stations nearby.

House Prices by Flat Type and Size



All HDB flats in my data set come with minimally a living room, kitchen, a toilet and a bedroom (2-Room flats). A 3-Room flat has an additional master bedroom with an ensuite toilet. 4 and 5-Room flats have one and two additional bedrooms respectively. Executive and Multi-Generation flats are

[8] https://www.mynicehome.gov.sg/hdb-how-to/buy-your-flat/mature-non-mature-estates-whats-the-difference/

[9] https://www.nlb.gov.sg/main/article-detail?cmsuuid=1dc1cdc8-1f9f-4a52-8237-71a768739ef7#

larger than standard flats, with the latter having more rooms to home families with three or more generations (grandparents, parents and children).

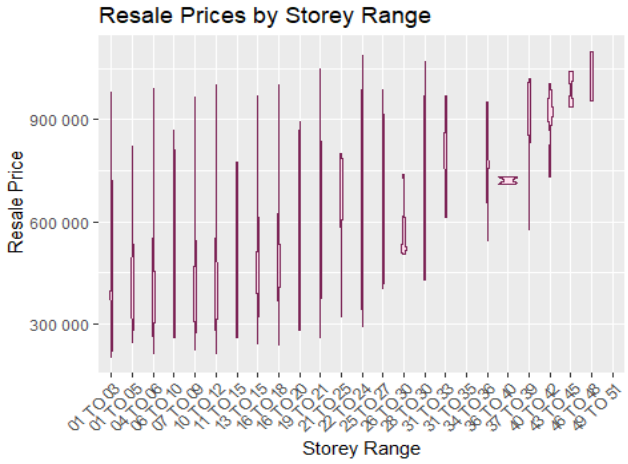
Naturally, there is positive relationship between the size of the flat and its resale price. Singaporeans are also willing to pay a premium for more space. As work from home becomes increasingly common after Covid-19, demand for bigger flats with more rooms to transform into home offices surged[10]. Coupled with Singaporean's ageing population and increased childcare costs, it is typical for grandparents to live with their children and grandchildren, contributing to the demand for bigger spaces.

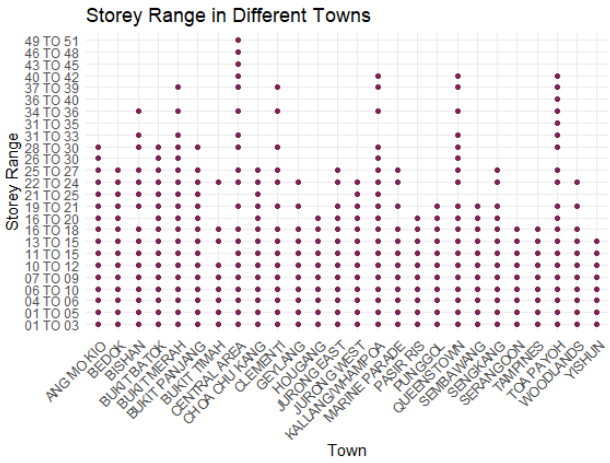
House Prices by Storey Range

Higher storey flats fetch a higher resale price thanks to the notion that higher levels offer greater privacy, unblocked views and are windier.

The plot on the right seems to affirm that. The lower bound for higher stories from floor 26 and higher is significantly higher.

However, this it should also be noted that more mature or central towns tend to have higher stories. The only town with storey range as high as 49 to 51 is Central Area. Prime locations such as Kallang, Queenstown, Toa Payoh and Bukit Merah make up the majority of towns with flats that offer such high floors. Part of the high resale price could be attributed to the premium of the town.

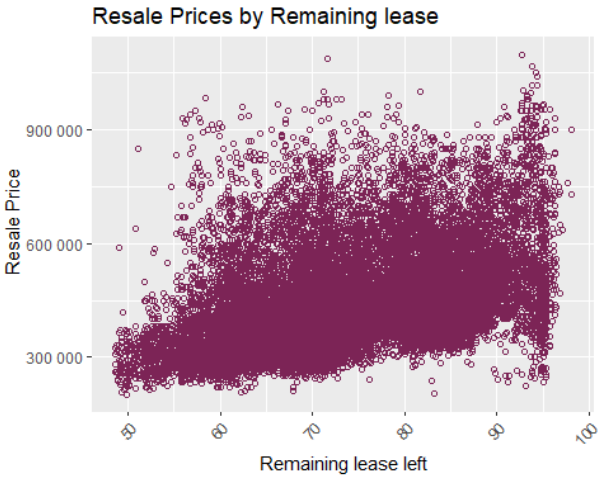




House Prices by Remaining Lease

At the end of the flat's 99-year lease, the flat must be returned to the state. This means that the value of the house should decrease when it approaches the end of its lease. The rationale behind such a policy is that 99 years is enough to last a persons' lifetime, after which the flat can be used by future generations[11].

From the graph, it seems that the resale price of the flat only starts to reflect its depreciated value when it has less than 55 years of lease remaining.



[10] https://www.channelnewsasia.com/commentary/bigger-flats-record-price-hdb-private-pandemic-whampoa-terrace-1935776

[11] https://www.channelnewsasia.com/singapore/ndr-2018-hdb-lease-99-years-flat-national-day-rally-804611

Modelling

I split the data into training and test with a 0.75 proportion. An empty linear regression and random forest model was created. For the random forest model, I set the importance as "permutation" as I had both categorical and continuous variables. Using the fit_xy function from parsnip, I fitted my data_train onto both models.

Results

```
## Residual standard error: 32160 on 14656 degrees of freedom
## Multiple R-squared:  0.9478, Adjusted R-squared:  0.9387
## F-statistic: 103.4 on 2576 and 14656 DF,  p-value: < 2.2e-16
```

For the linear regression model, r^2 and adjusted r^2 seems high, at 0.9478 and 0.9387 respectively. RSS explains variation not attributable to the relationship between the specified independent and dependent variables (Zach, 2021). A smaller RSS means the model is better at explaining the relationship.

```
## OOB prediction error (MSE):       1156867804
## R squared (OOB):                  0.931407
```

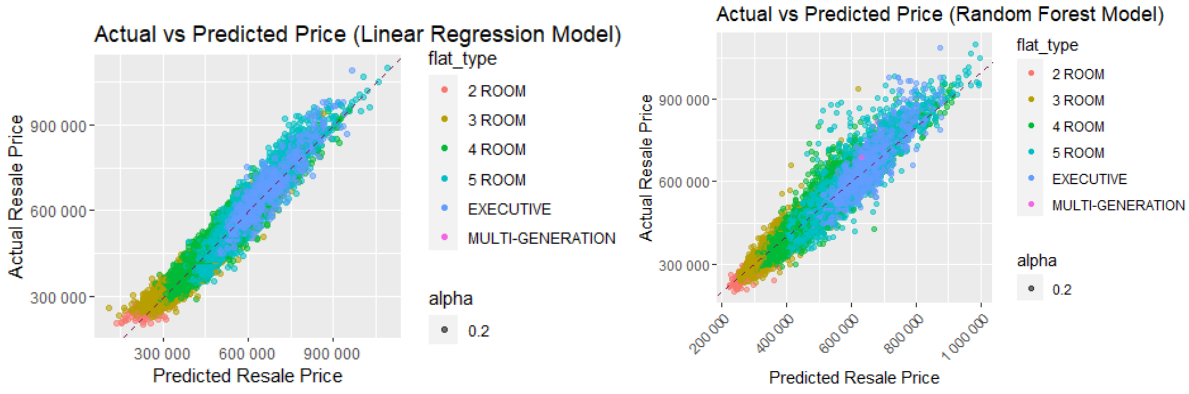
For the random forest, r squared (OOB) is 0.9323342. The OOB score is an estimate of the model's performance on unseen data, somewhat akin to a cross-validation score. It indicates how well the model generalizes to new, unseen data without the need for a separate validation set and is not directly comparable with the linear model's r^2 (Janitza & Hornung, 2018).

RMSE/MSE:

```
# rmse of linear regression model
Metrics::rmse(Lin_reg_fit$fit$fitted.values, data_train$resale_price)

## [1] 29656.5

# rmse of random forest model
sqrt(Rand_for_fit$fit$prediction.error)

## [1] 34012.76
```

Both models have high RMSE, meaning they are not accurate. The average price predicted is wrong by around $29,656 for the linear regression model, and $34,012 for the random forest model.

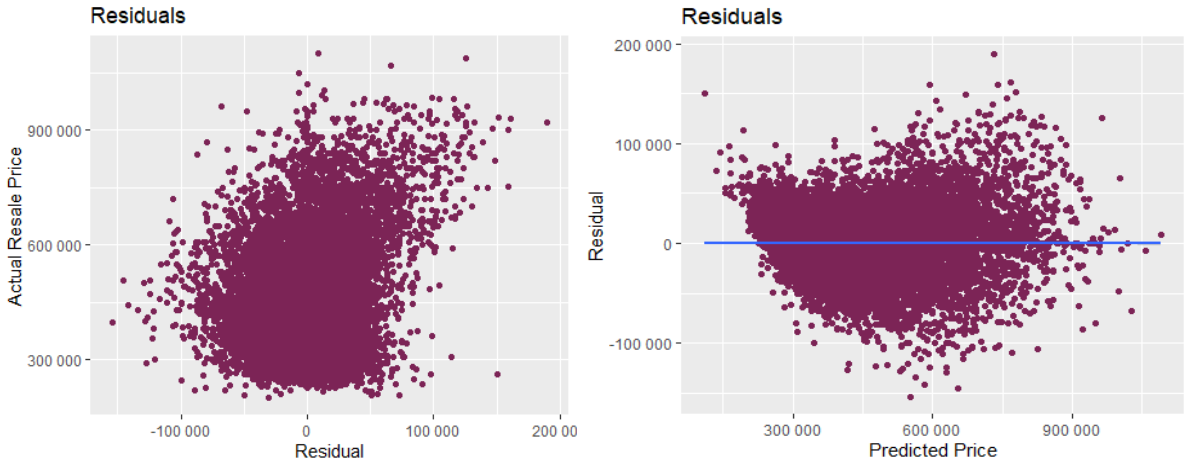
Actual vs predicted prices plot



I plotted the actual against predicted price and obtained a distinct linear relationship between the two for both models. Both models are accurately capturing the relationship between the features and target variables. But for the Random Forest Model, the plots around the median resale price spread out more. This suggests heteroscedasticity, meaning that the Random Forest Model's errors are not consistent across its range of predictions (Heteroscedasticity, 2022).

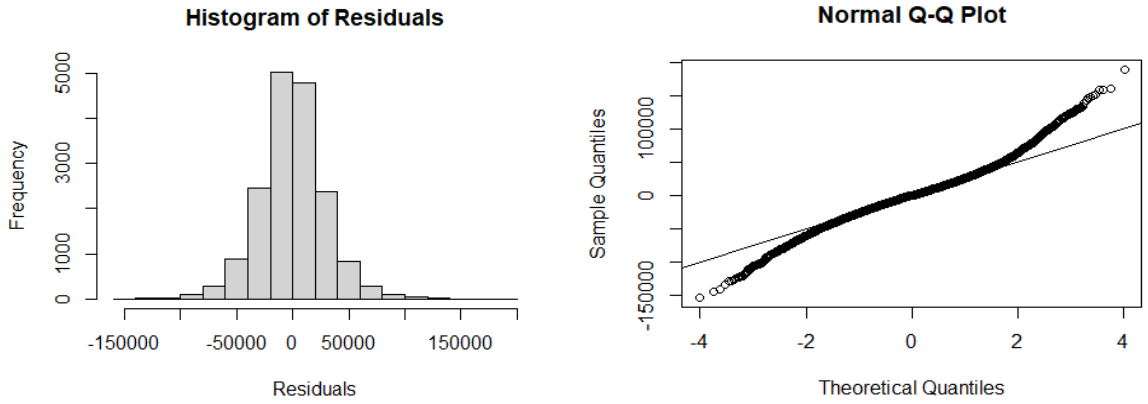
At higher resale prices, more of the points fall above the dotted line. At lower resale prices, more points fall below the dotted line. This suggests that the Random Forest model tends to overestimate at higher prices and underestimate at lower prices.

Residuals (Linear Regression)



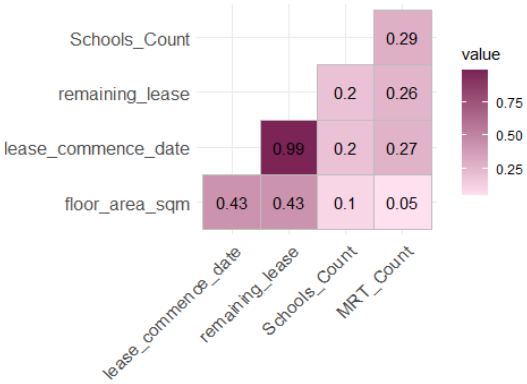
On the residual vs actual price plot on the left, the points appear randomly scattered around the zero line. There is a slight discernible tendency for higher resale prices to have positive residuals and lower resale prices to have negative residuals. The linear regression model might also overestimate higher prices and underestimate lower prices, but it is more accurate across the range of predictions than the random forest's.

The plot on the right shows that residuals appear to lack homoscedasticity. There is no discernible linearity, which shows that the variability of the residuals is not consistent throughout the range of predicted values.



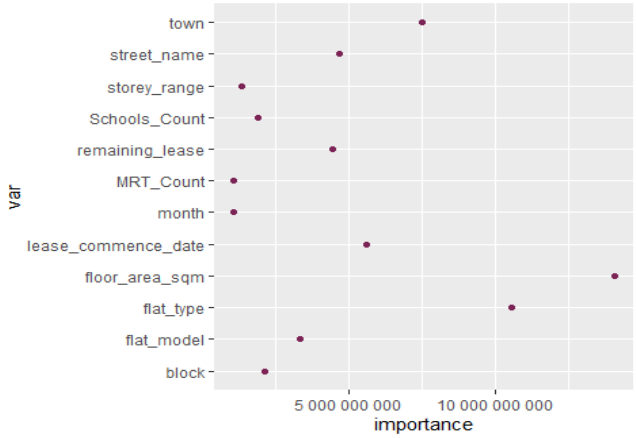
Residuals are normal except for extreme ends. This corresponds with the plots above, that the residuals do not have perfect homoscedasticity.

Multicollinearity (Linear Regression)



Another assumption of linear regression is there should be no multicollinearity, hence I will remove lease_commence_date as it is highly correlated with remaining_lease, since I calculated remaining lease from lease commence date (Kim, 2019).

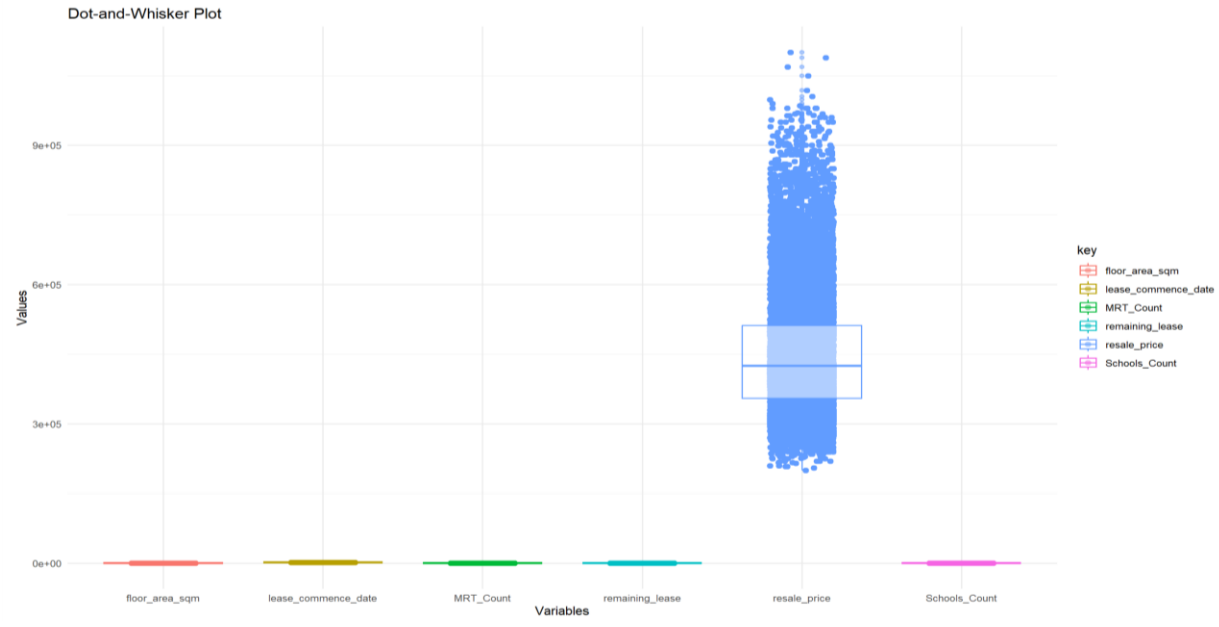
Variable Importance (Random Forest)



It seems like the number of MRT stations and the month the house was sold for contribute the least to the model, but I will still include them. As expected, the size of the flat plays the biggest role in increasing its resale price.

Model Tuning

Standardization



As the scale and range of my numerical variables differs drastically, I will apply standardization to all of them. This will ensure the ranges of all features are similar, and no feature disproportionately influences the model. There is no need to do this for tree-based models, so I will skip this for my random forest model[12].

[12] https://stats.stackexchange.com/questions/262895/scaling-normalization-not-need-for-tree-based-models

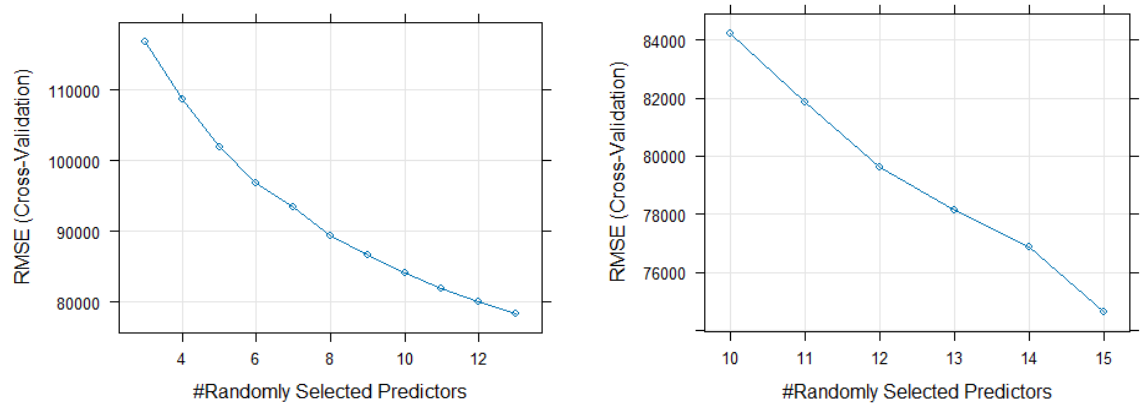
```
## Analysis of Variance Table
##
##   Res.Df        RSS Df   Sum of Sq  F Pr(>F)
## 1  14656 1.5157e+13
## 2  14656 8.9900e+02  0  1.5157e+13
## 3  14657 9.0000e+02 -1 -1.0000e+00  0      1
```

Model 1 is the original model while 2 is with standardised variables, and 3 is with standardised variables and without lease_commence_date.

RSS improved significantly from model 1 to 2. No p-value associated with F statistic as both models are not nested. RSS does not decrease much from model 2 to 3. Additionally, the significance of the F statistics between model 2 and 3 is low, hence there is no difference if I included lease_commence_date. I will however, choose model 3 since it is simpler.

Random Search: mtry

The random search for optimal mtry was done on a random forest object instead of a ranger object as it was not possible on the latter. The method was set as cross validation with 5 folds. I did a random search instead of a grid search and used a subset of the training data to shorten the long computational time (see Appendix C).



When I tested mtry 3 to 13, 13 gave the lowest RMSE. A second random search with mtry 10 to 15 yielded 15 as the optimal mtry. However, from the graph, the gradient starts to become gentler around 9 or 10. Beyond mtry = 10, there is diminishing returns as the decrease in RMSE gets smaller. So mtry was set to 10 to prevent the model from getting too memory and time intensive.

Assigning Variable Weights

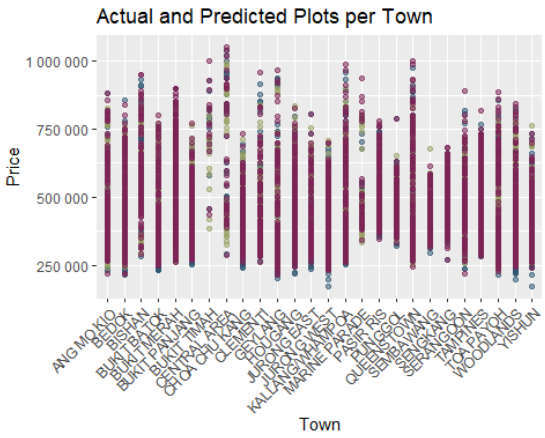
At each node splitting, a variable is selected from a random subset of variables (which is 10 variables as set earlier). However, since I know that some variables contribute greater, I could increase the probability that those variables are selected, thus increasing the accuracy of the model. Using the importance values obtained earlier, I set the probabilities of each variable using split.select.weights (see Appendix D).

```
Rand_for_fit3= ranger(resale_price ~  month + town + flat_type + block + street_name +
       storey_range + floor_area_sqm + flat_model + lease_commence_date +
       remaining_lease + Schools_Count + MRT_Count,
       data = data_train,
       importance = "permutation",
       split.select.weights = Importance_Var$weights, mtry = 10)
```

The OOB improved, so the model becomes less likely to overfit. It increased from 0.931407 to 0.9416355. RMSE also decreased from 34,012 to 31,299.

Visualize & Evaluate Output from Test Data

After the final tuning of both models, I fitted the data again. For the linear model, standardized inputs were used, hence the predicted prices were standardized as well. I reversed the standardization and put both model's predictions on the test data.




```
summary(Lin_reg_fit_test$fit)
```

```
## Residual standard error: 0.2475 on 3647 degrees of freedom
## Multiple R-squared:  0.9611, Adjusted R-squared:  0.9388
## F-statistic: 42.98 on 2098 and 3647 DF,  p-value: < 2.2e-16
```

```
Rand_for_fit_test
```

```
## OOB prediction error (MSE):       1793196602
## R squared (OOB):                  0.893225
```

The multiple r^2 of the linear model has increased slightly from 0.9478 to 0.9611. Adjusted r^2 remains relatively constant. However, since my tuned model has one less predictor, adjusted r^2 is more appropriate to compare (CFI Team, 2023). Hence, there is no increase in the variability of the resale price explained by the other predictors. The MSE of the random forest tree has also increased from 1,156,867,804 to 1,793,196,602. OOB decreased from 0.9416355 to 0.893225.

In the plot on the top right, the blue dots are the predicted prices by the linear regression model, green is by the random forest model, and red is the actual sales price. Visually, the better model appears to be the linear regression model as we see less blue than green spots.

RSME of Test Data Output

```
# rmse of linear regression model
Metrics::rmse(data_test$unstandardised_predictions_lr, data_test$resale_price)

## [1] 25547.7

# rmse of random forest model
sqrt(Rand_for_fit_test$prediction.error)

## [1] 42346.15
```

The linear regression model performs better as its RMSE is almost half of the random forest model's. The random forest performed worse on the new, test data, possibly because it was overfitted on the training data. Although the RMSE of the linear regression model is much lower, I think it is still too high to satisfactorily predict resale prices. A variance in $25,000 might only be approximately 5% of the house price, but for the average person, it is a large sum.

Discussion & Future Work

Summary of Results

This paper used Linear Regression and Random Forest Tree to model resale prices against multiple variables. Initial RMSE of the Linear Regression model was 29,656 and that of the Random Forest was 34,012. After scaling the variables and decreasing collinearity for the Linear Regression model,

RSS decreased from 1.5157e+13 to 8.99e+2, while r^2 does not change much. As for the Random Forest Tree model, after setting mtry to 10 and assigning probability weights for node splitting, RMSE decreased to 31,301 and OOB increased from 0.931407 to 0.9416355. However, upon testing on new data, the Random Forest Tree model performed poorer than the Linear Regression model, giving a RMSE of 42,346 compared to 25,547.

In conclusion, houses with high resale prices have the following attributes: residing in mature and central town locations like Central Area and Queenstown, bigger in size, has at least 60 years remaining in its lease. Other variables like storey range surprisingly does not affect the resale price of the flat as much. The number of MRT stations and schools nearby also plays a smaller role in the resale value of the flat. However, this could be due to limitations of my analysis.

Limitations & Further Research

This paper did not explore what the profitability of the houses were. Although factors leading to a higher resale price was explored, higher resale value flats were initially sold at a premium. My model can only predict the approximate price a HDB flat could fetch in the resale market. Further analysis on the profitability of HDB houses would provide more valuable insights for consumers.

Additionally, Singapore's government introduced new policies to cool the overheating property market over the years of my dataset range. One of which was the introduction of flat classifications by location. Prime Location Public Housing (PLH) Model imposes additional restrictions on the reselling and renting of flats marked as 'PLH'. These restrictions would influence resale prices, but they were not considered in this paper.

The added count of primary schools under 2km is not a good variable. In future work, assessing the popularity of the primary schools nearby and assigning an aggregated score to each HDB listing be a better predictor. The utility of having MRT stations nearby could also be better quantified by the distance between the flat and the nearest station. Other facilities such as hawker centres and supermarkets were also important but not considered in this paper. Exploring the relationships between the availability of these amenities could provide more comprehensive insights.

References

CFI Team. (2023, November 21). *Adjusted R-squared*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/adjusted-r-squared/

Heteroscedasticity. (2022, October 14). https://encyclopedia.pub/entry/28997

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PloS one*, *13*(8), e0201904. https://doi.org/10.1371/journal.pone.0201904

Kim J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, 72(6), 558–569. https://doi.org/10.4097/kja.19087

Zach. (2021, May 11). *How to interpret residual standard Error.* Statology. https://www.statology.org/how-to-interpret-residual-standard-error/

Appendix A

The following code was to change the full name of certain street names of the SgZipCodes data into the shortened version so that I could join both data frames by the address column.

```
full_name = c('ROAD', 'AVENUE', 'STREET', 'CRESCENT', 'CENTRAL', 'PLACE',
'BUKIT', 'DRIVE', 'NORTH', 'GARDENS', 'CLOSE', 'LORONG', 'JALAN',
'COMMONWEALTH', 'SAINT', 'UPPER', 'SOUTH', 'NORTH', 'HEIGHTS', 'PARK',
'TERRACE', 'TANJONG', 'KAMPONG', 'MARKET')
short_name= c('RD', 'AVE', 'ST', 'CRES', "CTRL", 'PL', 'BT', 'DR', 'NTH',
'GDNS', 'CL', 'LOR', 'JLN', 'C\'WEALTH', 'ST\\.', 'UPP', 'STH', 'NTH', 'HTS',
'PK' ,'TER', 'TG', 'KG', 'MKT')
Replace_list <- data.frame( full_name, short_name)

for (i in 1:nrow(Replace_list)) {
  SgZipCodes$road_name <- str_replace_all(SgZipCodes$road_name,
Replace_list[i,1], Replace_list[i,2])
}

SgZipCodes = SgZipCodes %>% mutate(address = paste0(blk_no, " ", road_name))
%>%
  select(-searchval, -postal, -building, -blk_no, -road_name) %>%
  distinct()
```

Appendix B

The HDB data was obtained from https://beta.data.gov.sg/datasets and searching 'Resale Flat Prices.' The three file names are:

1. Resale Flat Prices (Based on Registration Date), From Mar 2012 to Dec 2014
2. Resale Flat Prices (Based on Registration Date), From Jan 2015 to Dec 2016
3. Resale flat prices based on registration date from Jan-2017 onwards

The primary school data was obtained from a GitHub user, Hui Xiang Chua, from https://github.com/hxchua/datadoubleconfirm. It is called 'primaryschoolsg.csv'.

	Name	Type	GenderMix	Area	Zone	PostalCode	Latitude	Longitude	PlacestakenuptillPhase2B
1	Admiralty Primary School	Government	Mixed	Woodlands	North	738907	1.442700	103.7995	130
2	Ahmad Ibrahim Primary School	Government	Mixed	Yishun	North	768643	1.433300	103.8321	51
3	Ai Tong School	Government-aided, SAP	Mixed	Bishan	South	579646	1.360300	103.8321	302
4	Alexandra Primary School	Government	Mixed	Bukit Merah	South	159016	1.291300	103.8233	82
5	Anchor Green Primary School	Government	Mixed	Sengkang	North	544969	1.391300	103.8863	101

Likewise, the MRT station data was obtained from the same user. It is called 'mrtsg.csv'.

	OBJECTID	STN_NAME	STN_NO	X	Y	Latitude	Longitude	COLOR
1	101	HARBOURFRONT MRT STATION	CC29	26678.344	27555.06	1.265473	103.8214	YELLOW
2	101	HARBOURFRONT MRT STATION	NE1	26678.344	27555.06	1.265473	103.8214	PURPLE
3	189	TELOK BLANGAH MRT STATION	CC28	25376.847	28138.97	1.270753	103.8098	YELLOW
4	150	MARINA SOUTH PIER MRT STATION	NS28	31287.706	28203.49	1.271337	103.8629	RED
5	130	LABRADOR PARK MRT STATION	CC27	24619.058	28313.63	1.272333	103.8029	YELLOW

The zip/postal code data was obtained from Kaggle uploaded by MyleeSG, from https://www.kaggle.com/datasets/mylee2009/singapore-postal-code-mapper

	postal	latitude	longtitude	searchval	blk_no	road_name	building	address	postal.1
1	398614	1.312763	103.8835	# 1 LOFT	1	LORONG 24 GEYLANG	# 1 LOFT	1 LORONG 24 GEYLANG # 1 LOFT SINGAPORE 398614	398614
2	398721	1.312390	103.8815	# 1 SUITES	1	LORONG 20 GEYLANG	# 1 SUITES	1 LORONG 20 GEYLANG # 1 SUITES SINGAPORE 398721	398721
3	629875	1.309135	103.6795	1 BENOI ROAD SINGAPORE 629875	1	BENOI ROAD	NIL	1 BENOI ROAD SINGAPORE 629875	629875
4	439731	1.305466	103.8957	1 BOSCOMBE ROAD SINGAPORE 439731	1	BOSCOMBE ROAD	NIL	1 BOSCOMBE ROAD SINGAPORE 439731	439731
5	659592	1.344619	103.7498	1 BUKIT BATOK STREET 22 SINGAPORE 659592	1	BUKIT BATOK STREET 22	NIL	1 BUKIT BATOK STREET 22 SINGAPORE 659592	659592

Appendix C

I used code from this website to do the mtry random search: https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/. The first search was with the parameters .mtry = 3:13, and the second with parameters .mtry = 10:15.

```
set.seed(123)
control <- trainControl(method="cv", number=5, search="random")
tunegrid <- expand.grid(.mtry=c(3:13))
```

```
rf_gridsearch <- train(resale_price~month + town + flat_type + block +
street_name + storey_range + floor_area_sqm + flat_model +
lease_commence_date + remaining_lease + Schools_Count + MRT_Count,
                       data=data_tune, method="rf", tuneGrid=tunegrid,
trControl=control)

rf_gridsearch
```

```
## Random Forest
##
## 3000 samples
##   12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 2400, 2399, 2399, 2401, 2401
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    3    116850.79  0.6346724  88428.40
##    4    108635.81  0.6430975  81756.63
##    5    101856.47  0.6675918  76138.12
##    6     96761.26  0.6823658  71914.98
##    7     93439.59  0.6953529  69240.52
##    8     89371.87  0.7140201  66083.59
##    9     86694.07  0.7256057  63772.86
##   10     84047.02  0.7381405  61700.89
##   11     81883.97  0.7458511  59907.39
##   12     80108.35  0.7563428  58573.04
##   13     78343.89  0.7601119  57114.12
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 13.
```

```
rf_gridsearch2
```

```
## Random Forest
##
## 3000 samples
##   12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 2402, 2400, 2399, 2400, 2399
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   10    84211.87  0.7360643  61741.96
##   11    81864.27  0.7447461  59783.52
##   12    79634.13  0.7551835  58002.37
##   13    78157.67  0.7597282  56812.34
##   14    76870.23  0.7653282  55716.75
##   15    74647.01  0.7742841  53954.04
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 15.
```

Appendix D

The idea to assign probability to variables came from https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/. Although the usage of importance values from the first random forest model might be not scientifically accurate, it did appear to improve the model.