Canada MLS Text Data Analysis

Hyejin, Shim

September 14, 2020

1 Introduction

An analysis was conducted to identify what is the most meaningful keywords affecting customers in real estate transactions. The data was based on Canada MLS real estate information and analyzed and visualized using text mining techniques.

1.1 Canada MLS Data

Data obtained through the Multiple Listing Service (MLS) system, a site that collects all real estate sales, operated by the Canada Real Estate Association (CREA).

This data is real estate sales information observed over the range of 2013-05-14 to 2018-12-28 as of the date of the contract. The total number of records in datasets is 30,854 and the total number of variables representing information about the sale, such as regions, prices, and bedrooms, is 45.

The text analysis was performed by removing the missing value of the "remarksforclients" variable, detailed description of the customer. A total of 30,747 "remarksforclients" variable information was used for text analysis
The following table describes the variables in the MLS data set.

Table 1: Column of MLS Data

Variable name	Type	Definition
mlsnumber	chr	MLS number
countyorparish	chr	County or Parish
city	chr	City
postalcode	chr	Postal code
address	chr	Address
streetnumber	chr	Street number
unitnumber	chr	Unit number
pendingdate	chr	Pending date
closedate	chr	Close date
listingcontractdate	chr	Contract date
dom	int	Days On Market
cdom	int	Cumulative Days On Market
closeprice	num	Transaction price
currentprice	num	Current price
listprice	num	List price
originallistprice	num	Original list price
bedstotal	int	Number of beds
bathsfull	int	Number of large bathrooms
bathshalf	int	Number of small bathrooms
bathstotal	int	Total number of bathrooms
garageyn	chr	Garageyn
garagespacesnumber	num	Garage space number
lotfront	num	Lotfront
lotdepth	num	Lot depth
lotsize	chr	Lot size
yearbuilt	int	Built year
fireplacetype	chr	Fireplace type
waterfront	chr	Water front
pooltypes	chr	Pool types
latitude	num	Latitude
longitude	num	Longitude
occupancy	chr	Occupancy
sellername	chr	Seller name
photocount	int	Number of photos
remarksforbrokerages	chr	Remarks for brokerages
remarksforclients	chr	Remarks for clients

2 Text Preprocessing

Text data was converted to corpus, the basic structure of document management, in order to use "tm", a text mining package in the pre-processing. And the function of tm_clean was defined as follows to refine the Corpus according to the purpose of analysis. This function unifies in lowercase letters, removes sentence symbols and removes white spaces, removes unnecessary words, and finally removes numbers for each step.

```r
#----------------Cleaning----------------#
tm_clean = function(corpus){

  # step1. Change to lowercase.
  corpus = tm_map(corpus, tolower)
  cat("--------------------Print Step1--------------------",
      "\n[Step1.Change to lowercase]\n", corpus[[n]]$content)

  # step2. Remove punctuation.
  corpus = tm_map(corpus, removePunctuation)
  cat("\n--------------------Print Step2--------------------",
      "\n[Step2.Remove punctuation]\n", corpus[[n]]$content)

  # step3. Remove whitespace.
  corpus = tm_map(corpus, stripWhitespace)
  cat("\n--------------------Print Step3--------------------",
      "\n[Step3.Remove whitespace]\n", corpus[[n]]$content)

  # step4. Remove stopwords.
  corpus = tm_map(corpus, removeWords, stopwords(’english’))
  cat("\n--------------------Print Step4--------------------",
      "\n[Step4.Remove stopwords]\n", corpus[[n]]$content)

  # step5. Remove numbers (if numbers are unnecessary!)
  corpus = tm_map(corpus, removeNumbers)
  cat("\n--------------------Print Step5--------------------",
      "\n[Step5.Remove numbers]\n", corpus[[n]]$content, "\n")

  return(corpus)
}
```

After the text data was refined with the above function, stemming was performed to extract stem, which is a key part that contains the meaning of words. In the case of Stemming results, words that do not exist in the dictionary are often extracted, making it difficult to interpret them intuitively just by looking at the results of stemming. So I tried Lemmatization with features where the form of words is properly preserved.

```r
#----------------Normalization----------------#
## Stemming.(Using package of "SnowballC")
stem_remark = tm_map(remark, stemDocument)

## Lemmatization.(Using package of "textstem")
lem_remark = tm_map(remark, lemmatize_strings)
```

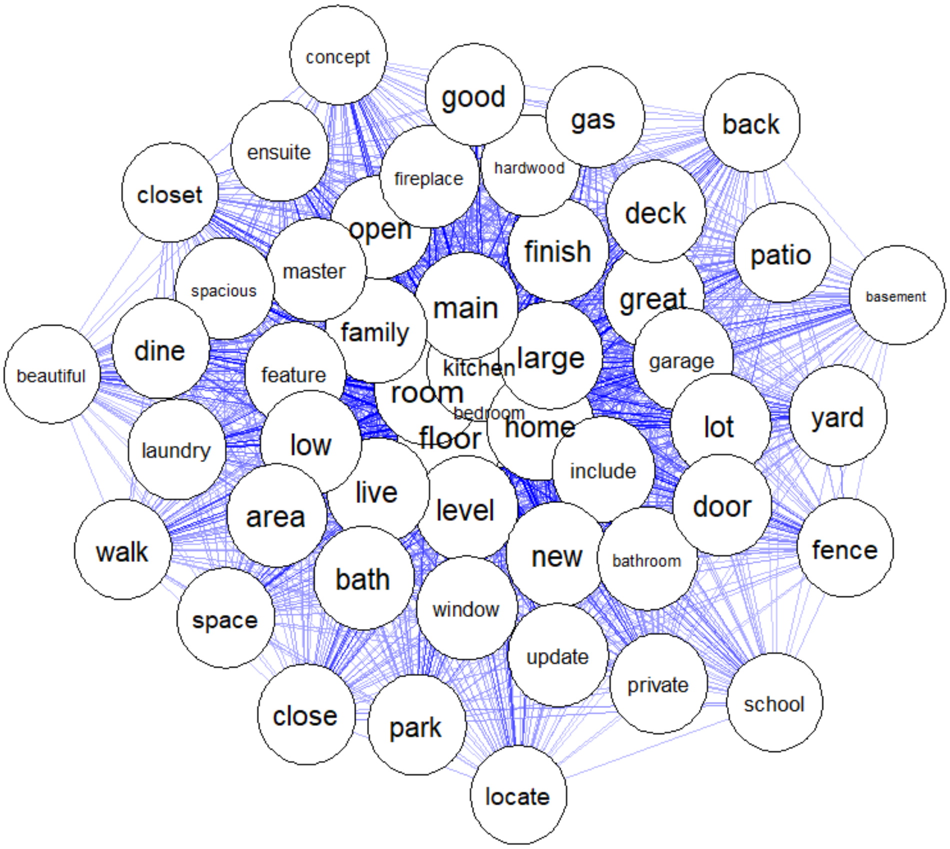
3 Frequency

To more intuitively interpret the results of the frequency analysis of words, results of Lemmatization was used. In order to find the frequency of words appearing in each document, the frequency of each word appearing in a number of documents was generated in a matrix (Term Document Matrix). When creating the matrix, the words were divided based on spaces, and the same repeated words in each document were counted only once.

As a result, Term Document Matrix consisted of a 30,749 documents and 24,506 words.

3.1 Co-occurrence Frequency

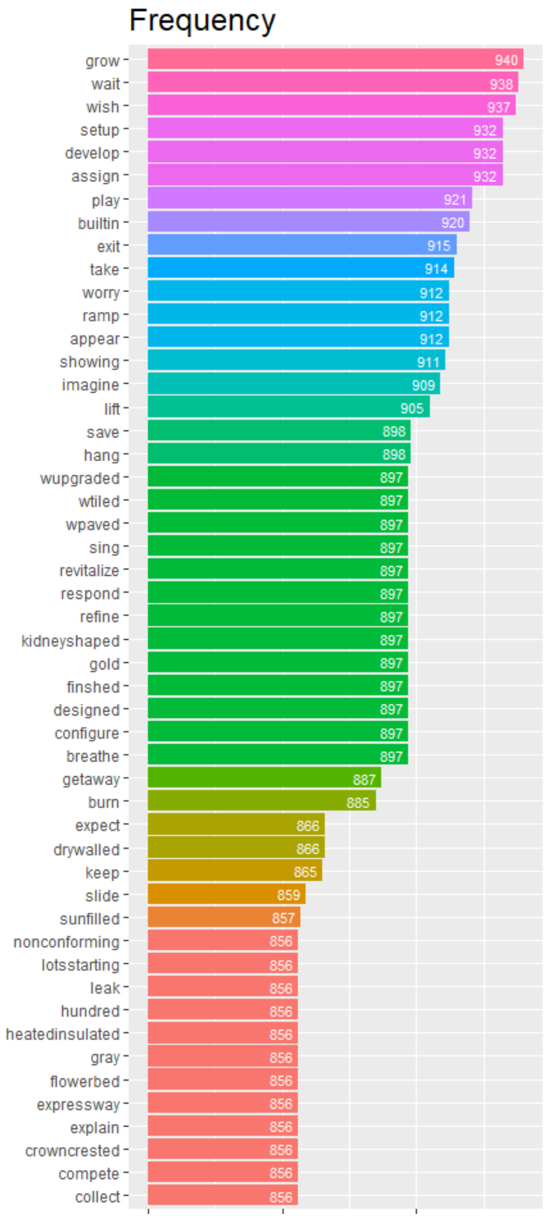
The co-occurrence frequency of the top 50 words was shown in a network graph. Co-occurrence is a word that appears simultaneously in one sentence and in a paragraph or text unit. This refers to semantic proximity and is used to find the collocation of words. That is why words related to each other are kept close together and words that are not.



3.2 Verb Frequency

This is the result of listing the top 50 verbs in order of frequency among a total of 2,012 verbs.

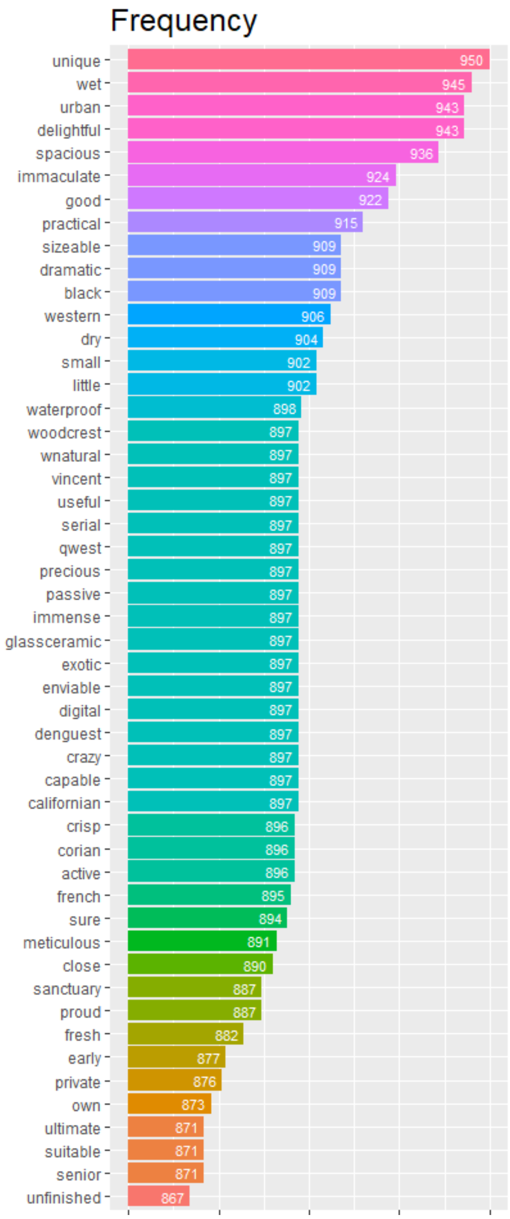
	Verb	Freq
1	grow	940
2	wait	938
3	wish	937
4	assign	932
5	develop	932
6	setup	932
7	play	921
8	builtin	920
9	exit	915
10	take	914
11	appear	912
12	ramp	912
13	worry	912
14	showing	911
15	imagine	909
16	lift	905
17	hang	898
18	save	898
19	breathe	897
20	configure	897
21	designed	897
22	finshed	897
23	gold	897
24	kidneyshaped	897
25	refine	897
26	respond	897
27	revitalize	897
28	sing	897
29	wpaved	897
30	wtiled	897
31	wupgraded	897
32	getaway	887
33	burn	885
34	drywalled	866
35	expect	866
36	keep	865
37	slide	859
38	sunfilled	857
39	collect	856
40	compete	856
41	crowncrested	856
42	explain	856
43	expressway	856
44	flowerbed	856
45	gray	856
46	heatedinsulated	856
47	hundred	856
48	leak	856
49	lotsstarting	856
50	nonconforming	856



3.3 Adjective Frequency

This is the result of listing the top 50 adjective in order of frequency among a total of 2,255 adjectives.

	Adjective	Freq
1	unique	950
2	wet	945
3	delightful	943
4	urban	943
5	spacious	936
6	immaculate	924
7	good	922
8	practical	915
9	black	909
10	dramatic	909
11	sizeable	909
12	western	906
13	dry	904
14	little	902
15	small	902
16	waterproof	898
17	californian	897
18	capable	897
19	crazy	897
20	denguest	897
21	digital	897
22	enviable	897
23	exotic	897
24	glassceramic	897
25	immense	897
26	passive	897
27	precious	897
28	qwest	897
29	serial	897
30	useful	897
31	vincent	897
32	wnatural	897
33	woodcrest	897
34	active	896
35	corian	896
36	crisp	896
37	french	895
38	sure	894
39	meticulous	891
40	close	890
41	proud	887
42	sanctuary	887
43	fresh	882
44	early	877
45	private	876
46	own	873
47	senior	871
48	suitable	871
49	ultimate	871
50	unfinished	867

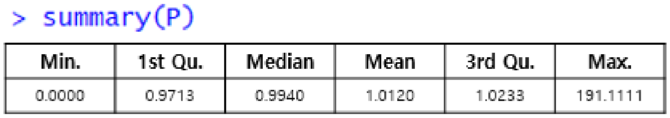


4 Keyword of Group

4.1 Comparison of keywords between groups

Based on the ratio(P) of the listing price and closing price, the group was divided into two groups to identify which key words affected each group's characteristics. In order to find the appropriate reference value for the price ratio (P), I first checked the descriptive statistics.

$$P = \frac{Closing\ Price}{listing\ Price}$$



```
> summary(P)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.9713	0.9940	1.0120	1.0233	191.1111

After dividing the groups by $P = 1$, It can see that the ratio of the two groups is properly divided by $H : L = 0.55 : 0.45$.

- if $P \geq 1$, then save data to "MLS_H". (14417 obs. of 47 variables)
- if $P < 1$, then save data to "MLS_L". (16321 obs. of 47 variables)

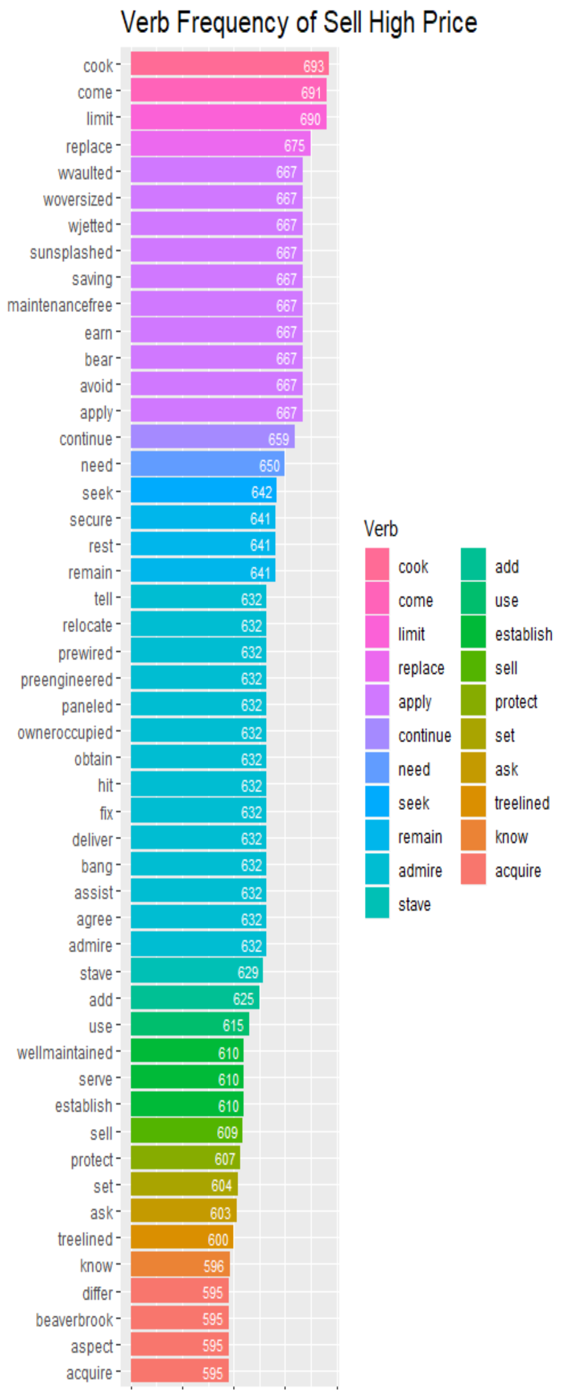
The following is the result of extracting word frequency by parts-of-speech (verb, adjective) from each of the two groups divided by Price Ratio (P) = 1. The frequency analysis of the words used the result of the Lemmatization that is simple to interpret. First of all, To find the frequency of words in each document, the words were divided by white space and the same repeated words in each document were counted only once. The total number of documents and words in both groups is shown below.

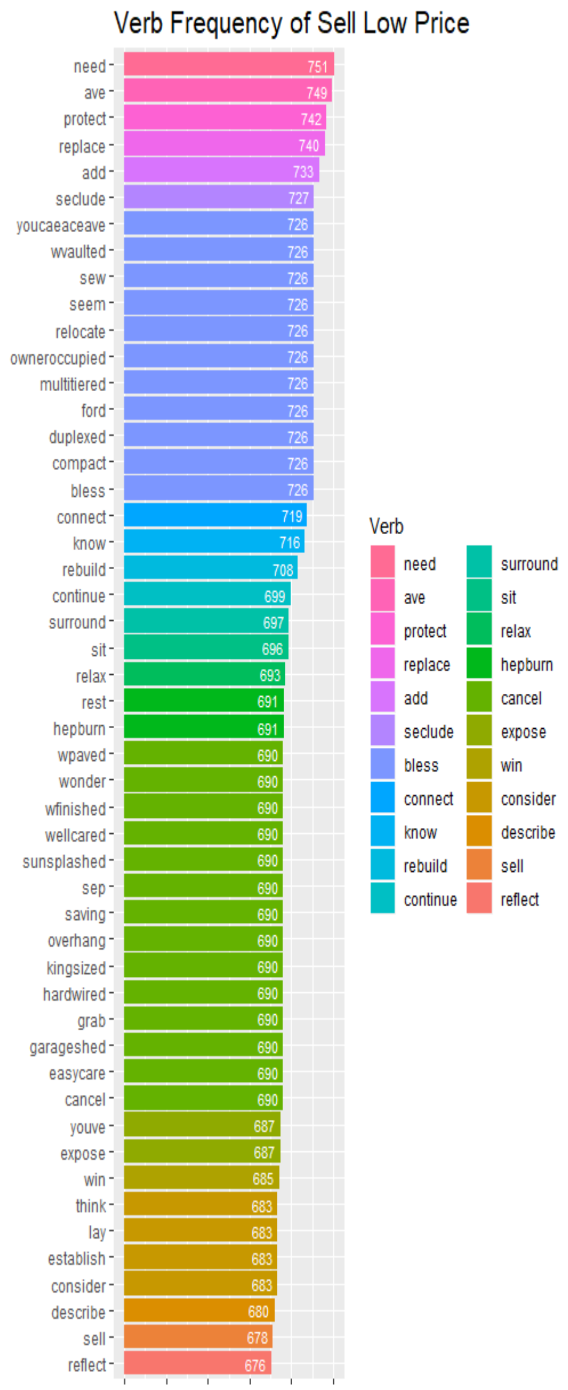
- Data "MLS_H" : consists of 14,417 documents and 15,492 words.
- Data "MLS_L" : consists of 16,321 documents and 17,569 words.

By using the above results, I redefined verbs and adjectives in each group, and wanted to see if there were differences between the two groups in the parts-of-speech words. The frequency of verbs and adjectives in each group is as follows.

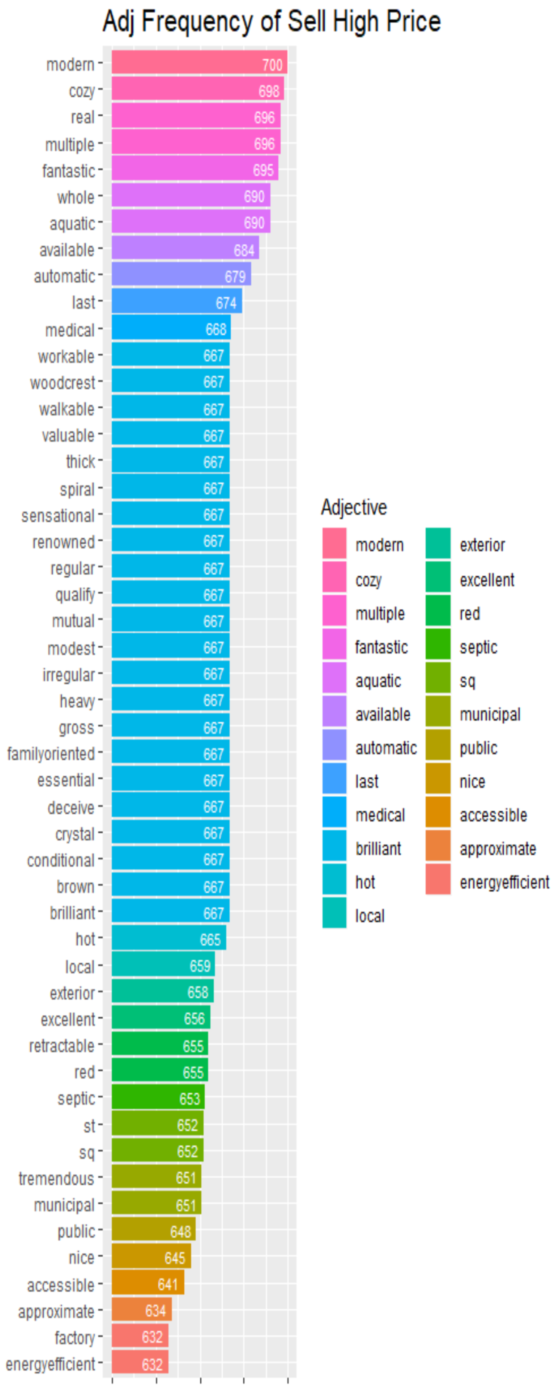
- Data "MLS_H" : 1,238 verbs / 1,468 adjectives
- Data "MLS_L" : 1,394 verbs / 1,705 adjectives

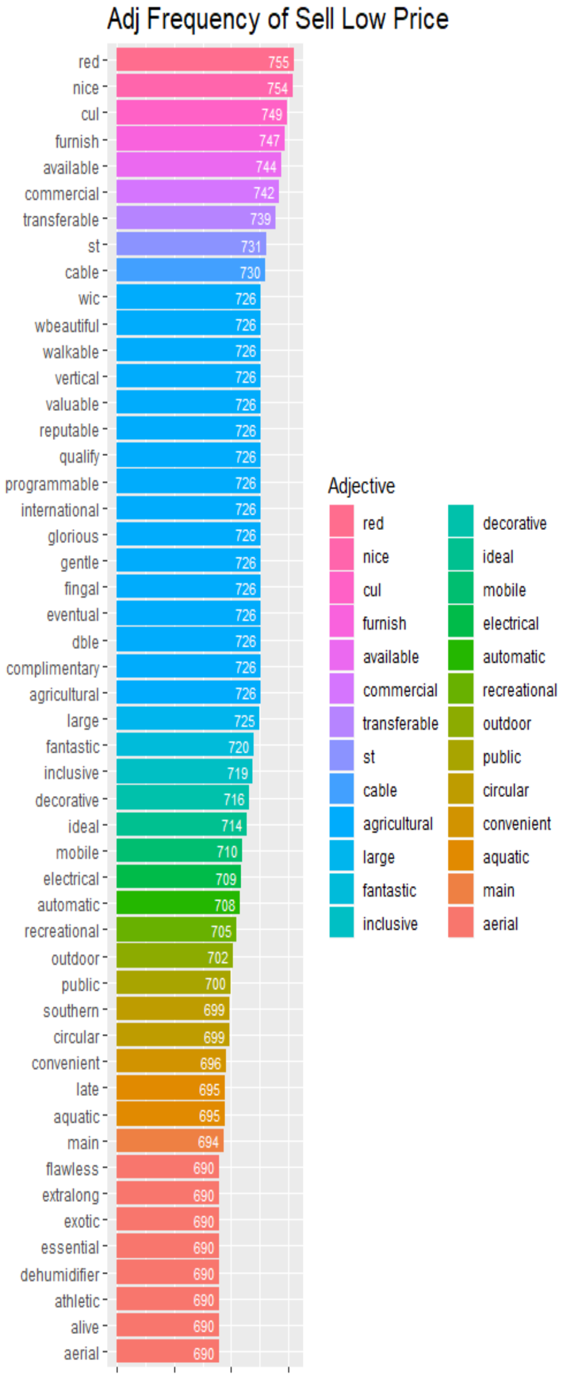
4.2 Verb


Verb Frequency of Sell High Price


Verb Frequency of Sell Low Price

4.3 Adjective





5 Words Difference

For those words with a frequency of less than 10 out of about 15,000 documents, I thought that it's not affected each group's characteristics. Therefore, I removed the words with a frequency of less than 10 and performed differential work on the two groups. The number of words excluded from each group is as follows.

- Verb : Group H - 706 of 1,238 (57%), Group L - 832 of 1,394 (60%)
- Adjective : Group H - 699 of 1,468 (48%), Group L - 847 of 1,705 (50%)

5.1 Verb Difference

The number of unique word sets for the verb is 118(10%) in group H and 148(11%) in group L.

Table 2: Unique word set of Group H

	Word	Freq
1	tell	632
2	set	604
3	ask	603
4	semisplit	595
5	chat	555
6	replete	555
7	bake	499
8	richfield	499
9	try	499
10	bevelled	420
11	grate	420
12	landscapedfenced	420
13	maint	420
14	meadowbrook	420
15	redbrick	420
16	warrantied	420
17	wattached	420
18	wtowering	420
19	awardwinning	333
20	bedroomscovered	333
21	belong	333
22	benshed	333
23	communitysouthgrove	333
24	conserve	333
25	dearborn	333

Table 3: Unique word set of Group L

	Word	Freq
1	seem	726
2	sep	690
3	wellcared	690
4	wpaved	690
5	etcwalk	657
6	nonconforming	657
7	realize	657
8	doored	597
9	homebased	597
10	replicate	597
11	wellconstructed	597
12	wellequipped	528
13	workshopbarn	528
14	wwoodburning	528
15	ashburn	452
16	eavestroughing	452
17	equiped	452
18	estatesized	452
19	freeze	452
20	louvered	452
21	repurposed	452
22	roomdinning	452
23	sauve	452
24	wamazing	452
25	wcoffered	452

5.2 Adjective Difference

The number of unique word sets for the adjective is 138(9%) in group H and 227(13%) in group L.

Table 4: Unique word set of Group H

	Word	Freq
1	brilliant	667
2	gross	667
3	regular	667
4	woodcrest	667
5	exterior	658
6	excellent	656
7	stable	632
8	illuminate	595
9	wclassic	555
10	wgorgeous	555
11	ascent	499
12	invisible	499
13	ongoing	499
14	sebastian	499
15	ceilingsconvenient	420
16	charismatic	420
17	elongate	420
18	extendable	420
19	exteriorinterior	420
20	initial	420
21	spinal	420
22	sweetbriar	420
23	undated	420
24	unify	420
25	variable	420
26	weekly	420
27	admiral	333
28	areaprofessional	333
29	artificial	333
30	beloved	333

Table 5: Unique word set of Group L

	Word	Freq
1	agricultural	726
2	glorious	726
3	wic(walk in closet)	726
4	exotic	690
5	phenomenal	690
6	precious	690
7	financial	657
8	multigenerational	657
9	occasional	657
10	restrictive	657
11	scandinavian	657
12	venetian	657
13	wextensive	657
14	american	597
15	economic	597
16	elaborate	597
17	inner	597
18	innovative	597
19	lucrative	597
20	official	597
21	opulent	597
22	unheard	597
23	visual	597
24	wireless	597
25	artistic	528
26	consistent	528
27	dental	528
28	enable	528
29	equestrian	528
30	equivalent	528

6 HPM with new text information

Summarize remarks in terms of continuous variable. and Apply a Hedonic Price Model(HPM) by adding the text variables as new predictor.

6.1 Readability Test

Readability is a method of evaluating the difficulty of text as a way to determine the level of text. There are various readability formulas for measuring readability. One of the most widely used formulas is the Flesch-Kincaid formula.

The most of studies about readability are based on words, sentence length, and frequency, which are factors of the text itself. Currently, there are over 100 readability formulas, but there are a few limited readability formulas due to their high utilization. What is the best formula of readability? Several studies have shown that there are no significant differences between the various readability test results and that instead of using a single measure, the average readability score is generated and used.

Based on the paper that examined the validation and accuracy of the readability formula, the readability score was calculated by selecting a numerical method from the formula provided by the readability function of R program.
* Harrison, C., Readability in the Classroom, Cambridge Educational, 1980.

The final data obtained is 53 variables and 30,745 observations. In the next step, I will use this data to run a regression analysis.

In the following formula, W is the number of words, St is the number of sentences, C is the number of characters, and Sy is the number of syllables. $W_{<3Sy}$ for words with less than three syllables, W^{1Sy} or N for words with exactly one syllable, W_{6C} for words with six or more letters, The number of words not in a specific word list is expressed as W_{-WL}.

6.1.1 ARI (Automated Readability Index)

$$ARI = 0.5 \times \frac{W}{St} + 4.71 \times \frac{C}{W} - 21.43$$

6.1.2 Coleman–Liau

$$CL = 141.8401 - 0.214590 \times \frac{100 \times C}{W} + 1.079812 \times \frac{100 \times S_t}{W}$$

6.1.3 Flesch Kincaid (Flesch-Kincaid Grade Level)

$$FK = 0.39 \times \frac{W}{St} + 11.8 \times \frac{Sy}{W} - 15.59$$

6.1.4 FOG (Frequency of Gobbledygook)

$$FOG = 0.4 \times \left(\frac{W}{St} + \frac{100 \times W_{3Sy}}{W} \right)$$

6.1.5 FORCAST

$$FORCAST = 20 - \frac{W^{1Sy} \times \frac{150}{W}}{10}$$

6.1.6 SMOG (Simple Measure of Gobbledygook)

$$SMOG = 1.043 \times \sqrt{W_{3Sy} \times \frac{30}{St}} + 3.1291$$

6.1.7 Linsear-Write

7 Regression Analysis

7.1 Correlation Analysis

A correlation analysis was conducted to determine whether the Readability Tests scores obtained in the previous section were linearly related to the closeprice variable. The following two tables show the Pearson's correlation coefficients obtained to determine the strength of the linear relationship between the closeprice and readability test score variables.

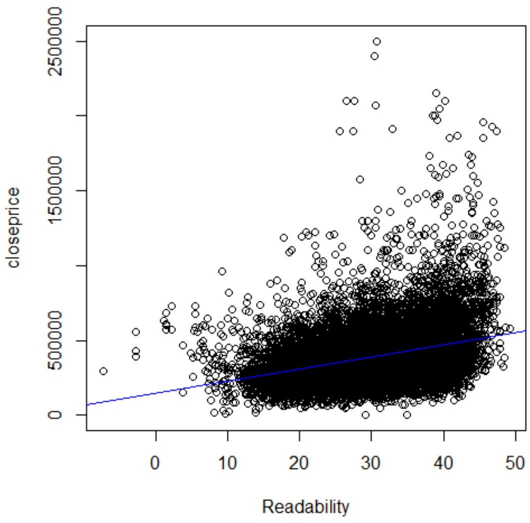
Table 6: Original data

correlation coefficient	closeprice
closeprice	1.00
Readability	0.16
ARI	0.17
ColemanLiau	0.16
FleschKincaid	0.15
FOG	0.12
FORCAST	0.07
LinsearWrite	0.14
SMOG	0.13

Table 7: Lemmatized process data

correlation coefficient	closeprice
closeprice	1.00
Readability	0.35
ARI	0.36
ColemanLiau	0.06
FleschKincaid	0.36
FOG	0.35
FORCAST	-0.01
LinsearWrite	0.36
SMOG	0.26

Comparing the correlation coefficient values of the variables in the above table, the readability scores obtained from the Lemmatized text have a higher correlation with the closeprice variable. The following is the output of the scatter plot of closeprice variables and the Readability variable, which represent the mean of the seven readability scores from the lemmatized text.



7.2 Regression Analysis

Based on the correlation analysis results, regression analysis was conducted to predict the mathematical relationship between variables with a certain pattern. The data used in the regression analysis consisted of 12 variables and 30,747 observations. The dependent variable used in the regression analysis was the "closeprice" variable and 9 other variables were used as the independent variable.

Table 8: Description of Regression Data

Variable name	Type	Definition
closeprice	num	Actual transaction price
listprice	num	listing price
garagespacesnumber	num	Number of garage spaces
photocount	num	Number of photos
dom	num	Days on Market
bedstotal	num	Total number of beds
bathstotal	num	Total number of baths
yearbuilt	num	Built year
Readability	num	Average of readability scores

The missing values for each variable in the data were as follows. After excluding these missing values, there were 17,490 observations of the final data used in the analysis.

	Variable	Number of NA
1	closeprice	11
2	listprice	0
3	garagespacesnumber	9,690
4	photocount	0
5	dom	83
6	bedstotal	0
7	bathstotal	0
8	yearbuilt	6,682
9	Readability	1

The following table shows the regression analysis results for each model.

- Model 1

$$(closeprice) = \beta_0 + \beta_1 \times (Readability) + e$$

- Model 2

  $closeprice \sim Readability + garagespacesnumber + photocount + dom + bedstotal + bathstotal + yearbuilt$

- Model 3

  $closeprice \sim Readability + garagespacesnumber + photocount + dom + bedstotal + bathstotal + yearbuilt + listprice$

Coefficients	Model1	Model2	Model3
(Intercept)	94859.2 (<2e-16 ***)	-153459.0 (< 2e-16 ***)	7835.9 (6.26e-09 ***)
Readability	9286.3 (<2e-16 ***)	3798.1 (< 2e-16 ***)	183.9 (5.85e-13 ***)
garagespacesnumber	-	93158.94 (< 2e-16 ***)	1288.7 (2.27e-04 ***)
photocount	-	3304.2 (< 2e-16 ***)	193.1 (5.85e-13 ***)
dom	-	299.7 (< 2e-16 ***)	-159.6 (< 2e-16 ***)
bedstotal	-	39849.0 (< 2e-16 ***)	853.4 (5.26e-05 ***)
bathstotal	-	35149.2 (< 2e-16 ***)	1977.1 (< 2e-16 ***)
yearbuilt	-	-24.64 (4.34e-14 ***)	-1.91 (6.68e-04 ***)
listprice	-	-	0.95
Adjusted R-squared	0.15	0.48	0.98

Table 9: Result of Regression analysis