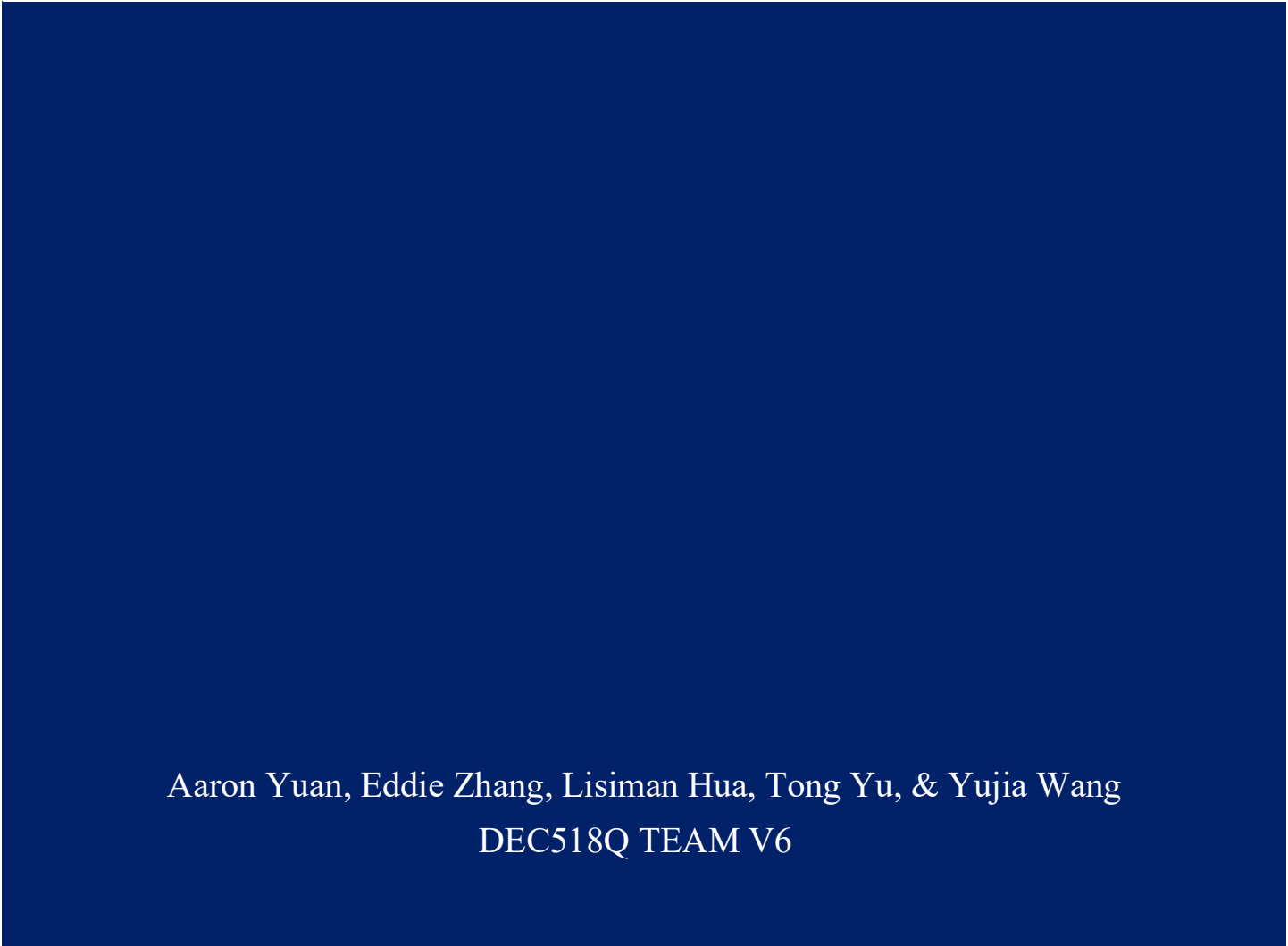




PREDICITING APARTMENT PRICES IN THE DAEGU HOUSING MARKET



Aaron Yuan, Eddie Zhang, Lisiman Hua, Tong Yu, & Yujia Wang
DEC518Q TEAM V6

Dataset Link

<https://www.kaggle.com/gunhee/koreahousedata>

Table of Contents

Business Understanding	1
Data Understanding	2
Data Preparation	2
Modeling	3
Evaluation	7
Deployment	10
Risks & Future Considerations	11

Business Understanding

Since the global financial crisis of 2008, the real estate market in South Korea has rebounded significantly. With housing prices skyrocketing in Seoul, the capital of South Korea, investors have turned to other cities. This has led to stable housing appreciation across major cities in South Korea. Daegu, the fourth largest city in South Korea, has been a top choice for individual buyers and investors. At the end of 2018, Daegu was the third hottest real estate market in South Korea.¹ In addition, the South Korean real estate market is different from the U.S. housing market. For example, high-rise apartments, buildings typically with 5 to 20 floors, are the most common form of housing in South Korea. As of 2018, such apartments made up 61.4% of registered housing.² Given the rising real estate prices, tightening mortgage policy³, and unique market characteristics, people want better information on apartments before buying.

Based on these needs, our objective is to better understand and predict apartment prices in Daegu, South Korea. A quick analysis of 5,891 apartment sales in Daegu showed a price range

¹ <https://www.globalpropertyguide.com/Asia/South-Korea/Price-History>

² <https://www.statista.com/statistics/877327/south-korea-residential-building-distribution-by-type/>

³ Same as 1st Footnote

from \$33K to \$586K. In addition, the dataset provided many numeric and categorical variables, but little initial insight was available on their effect on pricing. To start, our team analyzed the dataset to determine the key explanatory variables to pricing. Afterwards, we built an appropriate model using variable transformation, fitting, and residuals analysis. From our data mining solution, we can operate as an independent third-party agency that provides insights to sellers, buyers, and investors on optimal pricing and intrinsic market value. From there, these individuals can make well-informed decisions regarding the Daegu Housing Market.

Data Understanding

The dataset used for our analysis was uploaded by user Gunhee Park from kaggle.com, a public data mining resources website.⁴ This dataset contains detailed information on traded apartments in Daegu, South Korea from 2007 to 2017. The source of the dataset was the Official Korean Government Portal.⁵ Because it was prepared by Korean governmental agencies, the dataset conveniently does not contain many missing inputs. The dataset comprises of 5,891 observations with 30 variables (Exhibit 1). Six of these are categorical variables while the remaining 24 are continuous variables. The key dependent variable from this dataset is ‘SalePrice’. The remaining 29 variables describe apartments from one of three aspects: time attributes, in-house amenities, or nearby facilities and schools. To create a predictive pricing model, we need to determine the highest correlated variables to ‘SalePrice’ out of the 29.

Data Preparation

Although our data was quite clean, it still required some preparation work. First of all, this dataset contains many variables with confusing names, such as N_APT, N_Facilities(Total).

⁴ <https://www.kaggle.com/gunhee/koreahousedata>

⁵ <http://apis.data.go.kr>

We believe it is important to clearly describe the variable through its name. Therefore, we changed all the variable names to more comprehensible versions. In addition, the original class types for variables were numeric or character, which does not reflect reality. To make the data more accurate, the character type variables transformed into factor type variables.

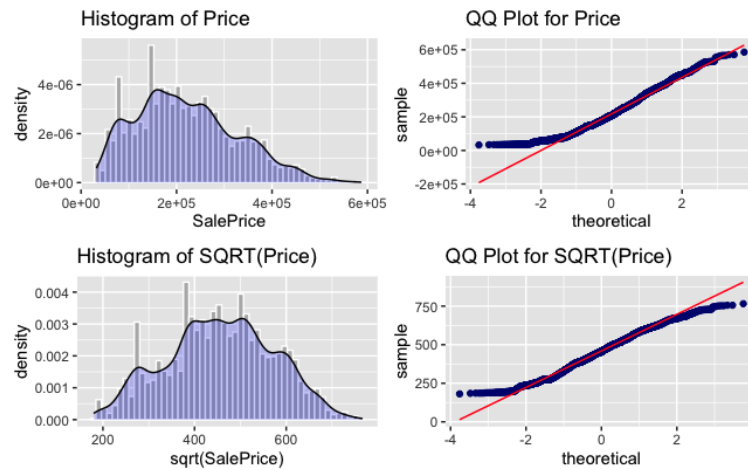
In addition, certain variables do not provide value for our analysis. In some categorical variables, such as HeatingType and AptManagerType, a main category occupies most of the observations (Exhibit 2). Given that the other categories are not significant enough, we excluded those variables. The TimeToSubway variable has a significant amount of 'NA' values, causing us to remove this bad data (Exhibit 3). Many of the facilities and school variables were also removed due to redundancy within the dataset.

Finally, we created 3 new variables for our data to potentially see certain relationships: Buildings, AgeWhenSold, and TotalParking. Buildings aggregated the number of public facilities by the apartment. AgeWhenSold took the difference between the YrSold and YrBuilt variables. TotalParking added the GroundParking and BasementParking variables together. After the process of adding and removing variables, our dataset now has 18 continuous variables and 3 categorical variables (Exhibit 4). This allowed us to run an initial correlation matrix on the continuous variables and potentially see any interesting price relationships (Exhibit 5).

Modeling

Based on our cleaned dataset, we chose a multiple linear regression as the most suitable model. Given that apartment price prediction is not a classification problem, any type of logistic regression would not work. Furthermore, a simple linear regression model with a single predictor would produce wildly inaccurate results. A scatter plot analysis of each dataset variable proved that no single variable had enough predictive power on its own. Therefore, we must identify,

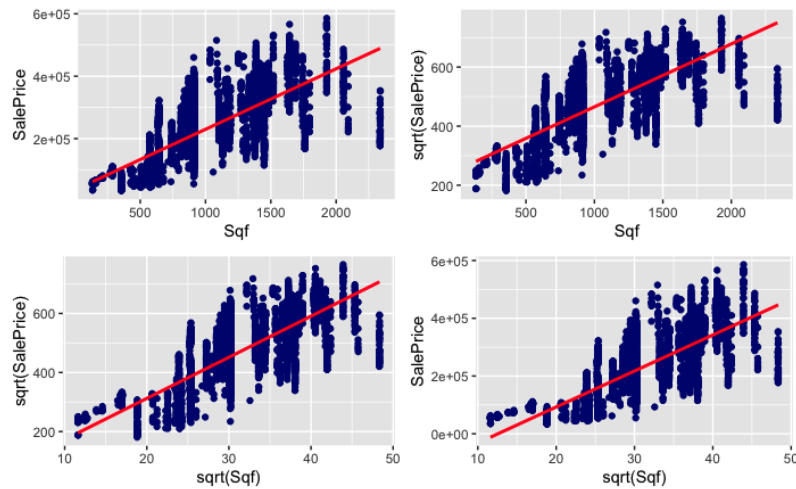
transform, and aggregate the most important variables in order to generate a model that achieves our ultimate goal.



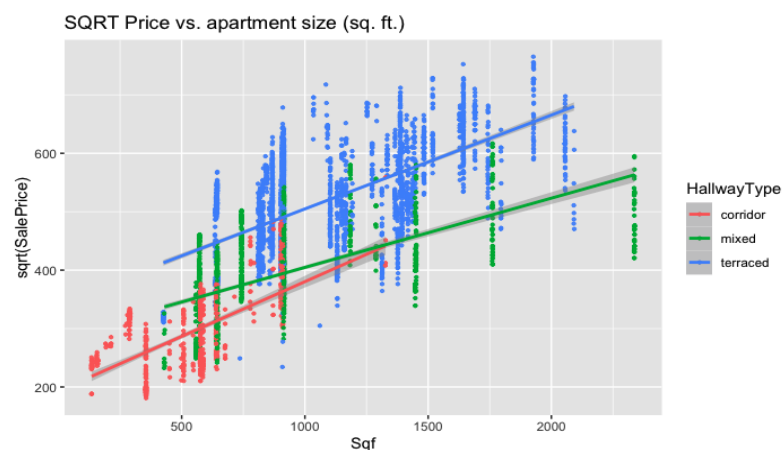
Before we searched for key independent variables, we first analyzed the characteristics of price, our dependent variable. As suggested by the top histogram and QQ plot, there is a right skew in the distribution of price, suggesting a lower density in high-end apartments coherent to common sense. A series of transformations on price were performed to correct the price distribution so that it becomes applicable for linear regression, which assumes normality and predictor-response linearity. Based on price distribution shaping and the original Box-Cox optimized calculation ($\lambda = 0.51$), a square root transformation would perform the best in achieving normality (Exhibit 6). This was confirmed by the bottom histogram, the bottom QQ plot, and the new Box-Cox calculation ($\lambda = 1.03$) (Exhibit 7).

After performing a square root transformation on price, we analyzed our first primary dependent variable, ApartmentSize. According to a correlation plot, ApartmentSize had the strongest positive correlation to SqrtPrice, which was 0.7 (Exhibit 8). However, the scatterplot below comparing ApartmentSize to SqrtPrice portrayed a curving right tail from the calculated line of best fit. This suggests that ApartmentSize would require a quadratic transformation to match the concave nature of the distribution. Because of this, our baseline regression model

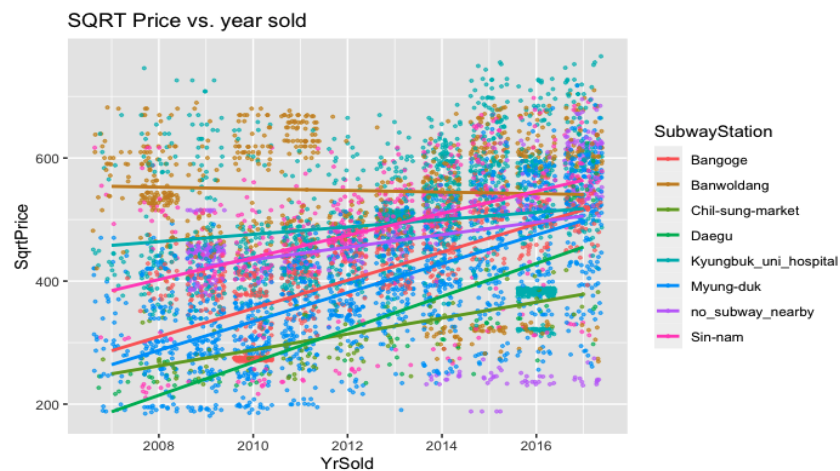
incorporated ApartmentSizeSquared and ApartmentSize as predictor variables. The initial raw model had a multiple R squared of 0.573, validating our choice of ApartmentSize as a variable.



In addition to ApartmentSize, we felt it was necessary to include a factored categorical variable in HallwayType as a multiplier. According to the scatter plot below, we can see that Apartment Size has strong interaction with Hallway type. Terraced apartments are more luxurious compared to mixed and corridor apartments of the same size. Meanwhile, mixed apartments have higher prices than corridor apartments. This makes sense because different hallway types can be related to other variables such as apartment age. Therefore, Hallway Type enhances the prediction of price based on ApartmentSize. When incorporating the HallwayType factored variable, the multiple R squared rises to 0.687 (Exhibit 9).



We also included the YrSold variable into our model based on its positive correlation of 0.4 with SqrtPrice (Exhibit 8). In addition, exploratory data analysis displayed a clear upward-sloping pattern. Intuitively, this makes sense given the price appreciation in the Korean housing market. For example, properties sold in 2016 would be more expensive than properties of the same qualities sold in 2012. Therefore, the YrSold variable acts as a time-series proxy for the linear regression.



However, further analysis demonstrated that the categorical variable, SubwayStation, interacted with the YrSold variable, as seen in the scatterplot above. Given that each category under SubwayStation consisted of a different trend line with a different intercept, it was highly imperative to use SubwayStation as a factored variable multiplied with YrSold. In essence, SubwayStation acts as a location proxy for the regression. Again, this makes sense as certain areas in Daegu are considered more luxurious than others. When including the YrSold and SubwayStation variables into our model, the multiple R-squared increases to 0.863 (Exhibit 10).

To improve our model, we included both the Amenities and the TotalFacilities variables. Amenities had a strong positive correlation of 0.5 with SqrtPrice and represented a numerical proxy for luxury (Exhibit 8). For example, an apartment value increases the more amenities are provided. Meanwhile, TotalFacilities had a strong negative correlation of -0.4 with SqrtPrice and

provided a different explanatory variable in our model (Exhibit 8). These additions yielded a multiple R-squared of 0.910 (Exhibit 11).

Finally, we added the SqrtFloors variable to the final model. Similar to price, we found out that the Floors variable needed transformation given its original histogram distribution and QQ plot. After performing a square root transformation, the distribution became more normal and closer to the 45 degree angle line in the respective visualizations (Exhibit 12). Furthermore, the Box-Cox optimized calculation suggested that it was the right transformation ($\lambda = 1.07$) (Exhibit 13). Now that we finally had all of our needed variables, our final regression produced a multiple R-squared of 0.914 and provided the smallest range in residuals (Exhibit 14). Our final model was achieved using the following code:

Lm(SqrtPrice~ApartmentSizeSquared+ApartmentSize*HallwayType+YrSold*SubwayStation+SqrtFloors+Ammenities+TotalFacilities, data = CleanApartmentData)

Evaluation

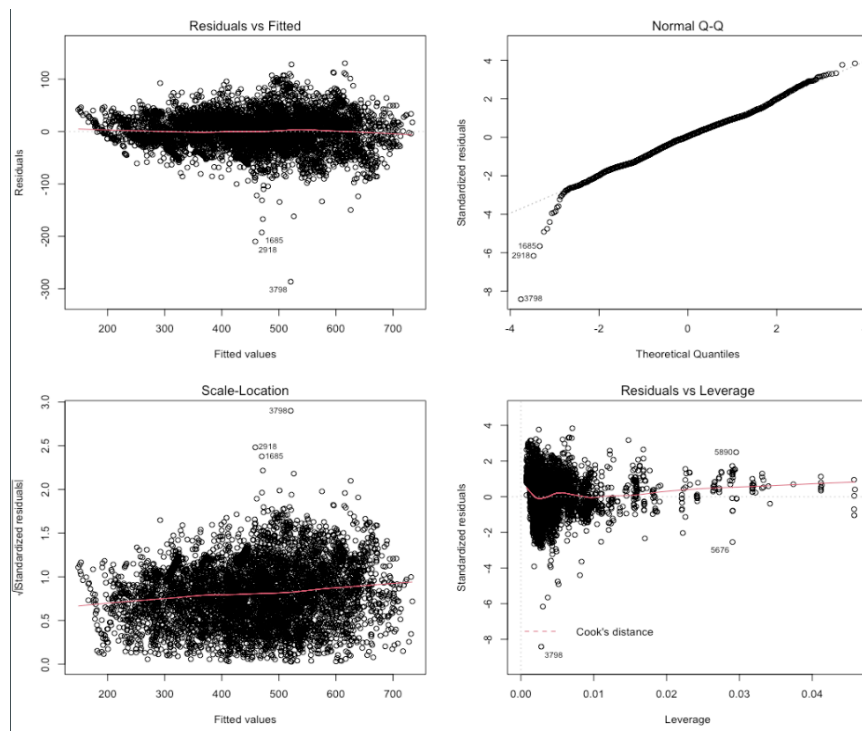
Given our final model, we found that all of the independent variables were statistically significant (p-value < 0.01). In addition, our model explains 91.4% of the variation in square-root price. Furthermore, our model returned a Residual Standard Error of 34.1 and an AIC value of 58,326, the lowest values compared to our other linear regressions. Finally, our model does not have overfitting problems because the number of observations (5,891) is significantly larger than the number of predictors (24). The key findings are as follow:

1. As expected, ApartmentSize, YrSold, SqrtFloors, and Amenities, positively correlates to price. For instance, for each unit increase in SqrtFloors, we expect price to increase by \$1.3225 on average. As the third largest metropolitan area by population (2.5 million), it is not suprising that apartment buyers would want to enjoy a peace of mind on higher floors.

2. We learned that the effect of ApartmentSize on price differs by HallwayType. The baseline HallwayType is Corridor, conveniently the cheapest one. The ApartmentSize effect is most pronounced for the subgroup of houses with Terraced hallway type, then followed by Mixed while controlling all other variables. Simply, for each square foot increase in ApartmentSize price will increase at a higher rate for Terraced apartments than Corridor apartments.
3. We learned that the price will be different when YrSold increases or decreases. In general, holding all else constant, every additional year in YrSold would lead to an average increase of \$434.31dollars in SalePrice. The effect of YrSold differs by SubwayStation. We know that the price of the apartments differs across locations, represented by the subway station nearby. Controlling all other variables, the YrSold effect is most pronounced for Banwoldang. We know that the most expensive apartments are near the Banwoldang. The rank-ordered effect of YrSold on SalePrice by subway station nearby in is Banwoldang, Kyungbuk-uni-hospital, Chil-sung-market, Sin-nam, Myung-duk, and lastly Daegu. For each year increase in YrSold and holding all other variables constant, we expect apartments near the baseline SubwayStationBangoge to increase SalePrice by \$434.31dollars, and apartments near SubwayStationBanwoldang to increase by \$29.38 dollars.
4. TotalFacilities negatively correlates with the price, with one more facility nearby the apartment, on average we expected the price to be \$21.16 lower. This is probably due to the additional noise and congestion, especially when it is a hospital or an elementary school.

When judging our model on simplicity, we believe that it would be easy to incorporate as the variables used in the linear regression are common information retrieved from the buying/selling process of property. Moreover, as mentioned before, each variable used in the model acted as a

proxy for characteristics that are intuitive to real estate pricing, such as luxury, location, timing, and congestion.



As for judging our model on effectiveness, we took a look at the residuals. The Residual vs Fitted plot above indicates that the mean of ε is very close to zero. Most of the observations' residuals congregate around the zero line, meaning the relationship is linear and few outliers exist in our model. Although Observation 1685, 2918, and 3798 are way off the theoretical line, our Normal Q-Q plot shows an approximate straight-line, indicating a normal distribution in the residuals. The Scale-Location plot also indicates an equal spread in variance across fitted SalePrice. Therefore, our model has met the homoscedasticity assumption of linear regression models. Finally, not all outliers are influential to our model fit as indicated by Residual vs. Leverage plot. All cases are within the Cook's distance line except Observation #3798. If we exclude observation #3798 from the regression model, the increase in R^2 is not significant.

After further investigation, the three indicated outliers -- Observation 1685, 2918, and 3798 -- were built in the 1980, indicating that our model underperforms for old apartments.

Regardless, customers interested in relatively new apartments (post-1980) can use our model with confidence. Finally, the quartiles of residuals in our model suggest that we can predict the price of an apartment within \$500 between the 1Q and 3Q, also known as the middle 50%.

Deployment

The primary use case scenario for this project is to help future real estate sellers in Daegu Metropolitan City to set optimal prices for their apartments based on the variables in our model. A secondary use case scenario for this project is to provide a “Buy or Wait” guide for on-sale apartments to local buyers in Daegu based on historical pricing data of properties with comparable attributes. Potential buyers can use our model as a reference and negotiate with sellers on a fair apartment price.

Our distribution will vary for different customers. Our website will provide the most comprehensive resources on Daegu Real Estate pricing. The website can show all available real estate on a 2-D map. If the buyers want to know the surroundings of their desired house, a 3-D satellite map will also be available. The location on a map is the most intuitive method for the buyers to choose their future home, making this a very useful feature. Moreover, a mobile app will be available for all types of customers. On mobile, customers can sort by independent variables (e.g. HallwayType, SubwayStation, etc.) to display their desired houses.

Our tools and service will be partially free for the sellers since the sellers rely on our business model to set the price. If they make a deal using our models, we will collect a commission fee from them. The individual buyers will pay a one-time fee for each real estate that they are interested in.

Risks and Future Considerations

The ethical issues that we face are the loyalty of the sellers. We have low control over the sellers. If the sellers decide to make a deal without telling us, we will suffer from the loss of commission fee. We must ensure that our information is seen as worthy for both the seller and buyer. Outside of that, real estate does not carry too many sensitive ethical considerations.

Furthermore, the risk of our business model is relatively low. We do not need to invest in real estate or physical resources. The only issue for our business model is we need to make sure the information on real estate is exclusive, authentic, and reliable. It is crucial that our data is up-to-date and quick-response to the market change.

Finally, when evaluating the integrity of the data we analyzed, we noticed that our dataset contains apartment trading information from only one of the seven districts in Daegu: the Buk District. It is only sufficient for us to focus our business inside the specific district. This district is less than 100km² thus our business can be limited. Another crucial point is that our dataset contains only information from 2007 to 2017. In an industry that can change abruptly, it is better for us to find more updated data if we want more accurate prediction. In the future, we can further expand our service by:

1. Adding time-series analysis to suggest how the pricing is changing overtime
2. Retrieve more timely information to consistently update our model to changing preferences
3. Retrieve a more updated dataset to include all the other districts in Daegu and compare pricing across the districts.

Appendix

Exhibit 1 [Table of the Original Variables from the Daegu Apartment Pricing Dataset]

#	Variable	Type	Content
1	SalePrice	integer	Price in US dollar (target feature)
2	YearBuilt	integer	Year apartment built
3	YrSold	integer	Year apartment sold
4	MonthSold	integer	Month apartment sold
5	Size.sqf.	integer	Size of apartment in square feet
6	Floor	integer	Floor level apartment located
7	HallwayType	character	Nominal categorical. 3 unique values: corridor, terraced, mixed
8	HeatingType	character	Nominal categorical. 2 unique values: individual_heating, central_heating
9	AptManageType	character	Nominal categorical. Management style of the apartment. 2 unique values: management_in_trust, self_management
10	N_Parkinglot.Ground.	numeric	Number of parking lot on ground
11	N_Parkinglot.Basement.	numeric	Number of parking lot in baseent
12	TimeToBusStop	character	Ordinal categorical. 3 unique values: 0~5min, 5~10min, 10~15min
13	TimeToSubway	character	Ordinal categorical. 5 unique values: 0-5min, 10min~15min, 15min~20min, 5min~10min, no_bus_stop_nearby
14	N_APT	numeric	Number of apartment building in the apartment complex
15	N_manager	numeric	Number of staff including manager, security, cleaner, etc
16	N_elevators	numeric	Number of elevator in the apartment complex

17	SubwayStation	character	Nominal categorical. Names of subway stations nearby. 8 unique values
18	N_FacilitiesNearBy.PublicOffice.	numeric	Number of public offices nearby
19	N_FacilitiesNearBy.Hospital.	integer	Number of hospitals nearby
20	N_FacilitiesNearBy.Dpartmentstore.	numeric	Number of department stores nearby
21	N_FacilitiesNearBy.Mall.	numeric	Number of malls nearby apartment
22	N_FacilitiesNearBy.ETC.	numeric	Number of facilities like hotels, special school, etc nearby
23	N_FacilitiesNearBy.Park.	numeric	Number of parks nearby
24	N_SchoolNearBy.Elementary.	numeric	Number of elementary schools nearby
25	N_SchoolNearBy.Middle.	numeric	Number of middle schools nearby
26	N_SchoolNearBy.High.	numeric	Number of high schools nearby
27	N_SchoolNearBy.University.	numeric	Number of universities nearby
28	N_FacilitiesInApt	integer	Number of in-region facilities like pool, gym, lounge
29	N_FacilitiesNearBy.Total.	numeric	Total number of nearby facilities = SUM(18:23)
30	N_SchoolNearBy.Total.	numeric	Total number of schools nearby = SUM(24:27)

Exhibit 2 [Summary of HeatingType and AptManageType Variables]

```
> summary(ApartmentData$HeatingType)
central_heating individual_heating
          300             5591
> summary(ApartmentData$AptManageType)
management_in_trust self_management
          5542             349
```

Exhibit 3 [Summary of TimeToSubway Variable]

```
summary(ApartmentData$TimeToSubway)
0~5min      5min~10min      10min~15min      15min~20min no_bus_stop_nearby
      0             1135             806             953             238
NA's
2759
```

Exhibit 4 [Summary of Variables in CleanApartmentData]

SqrtPrice	YrBuilt	YrSold	AgeWhenSold	ApartmentSize	Floors
Min. :181.0	Min. :1978	Min. :2007	Min. : 0.000	Min. : 135.0	Min. : 1.00
1st Qu.:379.8	1st Qu.:1993	1st Qu.:2010	1st Qu.: 3.000	1st Qu.: 644.0	1st Qu.: 6.00
Median :456.0	Median :2006	Median :2013	Median : 7.000	Median : 910.0	Median :11.00
Mean :455.7	Mean :2003	Mean :2013	Mean : 9.724	Mean : 955.6	Mean :12.03
3rd Qu.:539.6	3rd Qu.:2008	3rd Qu.:2015	3rd Qu.:16.000	3rd Qu.:1149.0	3rd Qu.:17.00
Max. :765.4	Max. :2015	Max. :2017	Max. :39.000	Max. :2337.0	Max. :43.00

HallwayType	Managers	Elevators	TotalApts	Amenities	TimeToBusStop
corridor: 637	Min. : 1.00	Min. : 0.00	Min. : 1.000	Min. : 1.00	0~5min :4509
mixed :1690	1st Qu.: 5.00	1st Qu.: 5.00	1st Qu.: 3.000	1st Qu.: 4.00	5min~10min :1327
terraced:3564	Median : 6.00	Median :11.00	Median : 7.000	Median : 5.00	10min~15min: 55
	Mean : 6.31	Mean :11.15	Mean : 5.614	Mean : 5.81	
	3rd Qu.: 8.00	3rd Qu.:16.00	3rd Qu.: 8.000	3rd Qu.: 7.00	
	Max. :14.00	Max. :27.00	Max. :13.000	Max. :10.00	

SubwayStation	Parks	DepStores	Buildings	TotalFacilities
Kyungbuk_uni_hospital:1644	Min. :0.0000	Min. :0.0000	Min. : 0.000	Min. : 0.000
Myung-duk :1507	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.: 6.000	1st Qu.: 8.000
Banwoldang : 748	Median :1.0000	Median :1.0000	Median : 7.000	Median : 9.000
Bangoge : 737	Mean :0.6542	Mean :0.8963	Mean : 8.033	Mean : 9.871
Sin-nam : 651	3rd Qu.:1.0000	3rd Qu.:2.0000	3rd Qu.:12.000	3rd Qu.:13.000
no_subway_nearby : 404	Max. :2.0000	Max. :2.0000	Max. :14.000	Max. :16.000
(Other) : 200				

TotalSchools	GroundParking	BasementParking	TotalParking
Min. : 0.00	Min. : 0.0	Min. : 0.0	Min. : 87.0
1st Qu.: 7.00	1st Qu.: 11.0	1st Qu.: 184.0	1st Qu.: 304.0
Median :10.00	Median :100.0	Median : 536.0	Median : 865.0
Mean :10.86	Mean :195.9	Mean : 570.8	Mean : 766.6
3rd Qu.:15.00	3rd Qu.:249.0	3rd Qu.: 798.0	3rd Qu.:1059.0
Max. :17.00	Max. :713.0	Max. :1321.0	Max. :1496.0

Exhibit 5 [Correlation Matrix Compared To Price]

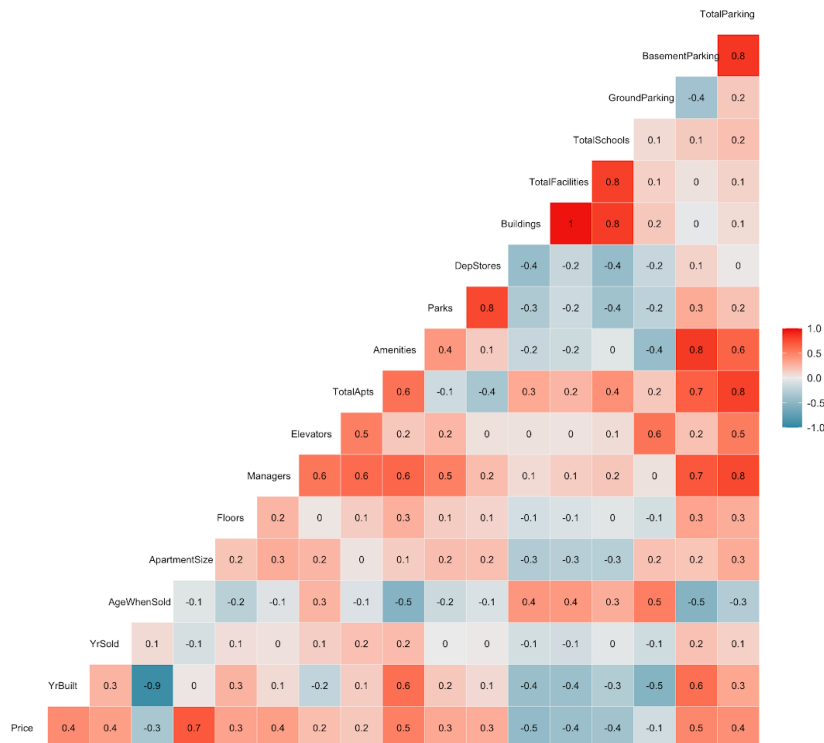


Exhibit 6 [Box Cox of Price Pre-Square-Root Transformation]

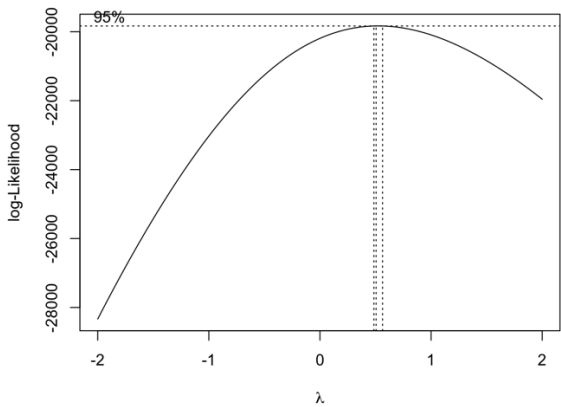


Exhibit 7 [Box Cox of Price Post-Square-Root Transformation]

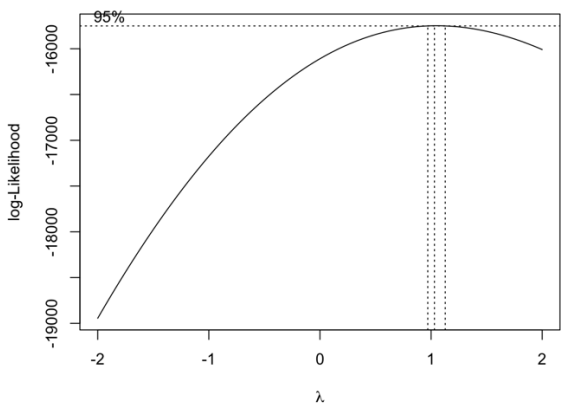


Exhibit 8 [Correlation Matrix Compared to SqrtPrice]

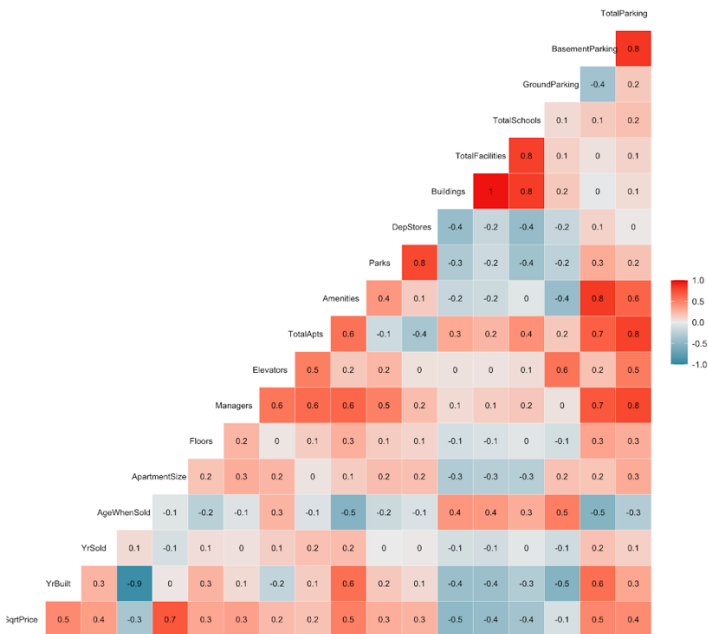


Exhibit 9 [Regression With ApartmentSize & HallwayType]

```
Call:
lm(formula = SqrtPrice ~ ApartmentSize * HallwayType + ApartmentSizeSquared,
    data = CleanApartmentData)

Residuals:
    Min       1Q   Median       3Q      Max
-259.342  -52.021   -1.992    49.341   190.037

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.721e+02  6.721e+00  25.607 < 2e-16 ***
ApartmentSize    2.762e-01  1.328e-02  20.790 < 2e-16 ***
HallwayTypemixed  9.155e+00  9.631e+00   0.951  0.34185
HallwayTypeterraced  6.582e+01  9.531e+00   6.905 5.53e-12 ***
ApartmentSizeSquared -7.961e-05  5.469e-06 -14.558 < 2e-16 ***
ApartmentSize:HallwayTypemixed  4.520e-02  1.480e-02   3.053 0.00227 **
ApartmentSize:HallwayTypeterraced  7.721e-02  1.416e-02   5.452 5.19e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.07 on 5884 degrees of freedom
Multiple R-squared:  0.6874,    Adjusted R-squared:  0.6871
F-statistic: 2157 on 6 and 5884 DF,  p-value: < 2.2e-16

> AIC(regression1a)
[1] 65922.63
```

Exhibit 10 [Regression With ApartmentSize, HallwayType, YrSold, & SubwayStation]

```
Call:
lm(formula = SqrtPrice ~ ApartmentSize * HallwayType + ApartmentSizeSquared +
    YrSold * SubwayStation, data = CleanApartmentData)

Residuals:
    Min       1Q   Median       3Q      Max
-273.855  -26.347    3.377    28.475   184.558

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.195e+04  1.248e+03 -33.615 < 2e-16 ***
ApartmentSize    3.718e-01  9.187e-03  40.471 < 2e-16 ***
HallwayTypemixed  1.578e+01  7.069e+00   2.232  0.0257 *
HallwayTypeterraced  4.389e+01  6.945e+00   6.320 2.81e-10 ***
ApartmentSizeSquared -1.025e-04  3.781e-06 -27.119 < 2e-16 ***
YrSold          2.092e+01  6.202e-01  33.738 < 2e-16 ***
SubwayStationBanwoldang  1.827e+04  1.625e+03  11.243 < 2e-16 ***
SubwayStationChil-sung-market  1.386e+04  3.380e+03   4.100 4.19e-05 ***
SubwayStationDaegu    5.008e+03  2.904e+03   1.725  0.0846 .
SubwayStationKyungbuk_uni_hospital  1.595e+04  1.505e+03  10.603 < 2e-16 ***
SubwayStationMyung-duk  5.873e+03  1.476e+03   3.979 7.02e-05 ***
SubwayStationno_subway_nearby -3.929e+03  1.816e+03  -2.163  0.0306 *
SubwayStationSin-nam  7.909e+03  1.835e+03   4.310 1.66e-05 ***
ApartmentSize:HallwayTypemixed  1.443e-02  1.046e-02   1.380  0.1677
ApartmentSize:HallwayTypeterraced  6.350e-02  1.012e-02   6.277 3.70e-10 ***
YrSold:SubwayStationBanwoldang -9.091e+00  8.077e-01 -11.255 < 2e-16 ***
YrSold:SubwayStationChil-sung-market -6.918e+00  1.680e+00 -4.118 3.88e-05 ***
YrSold:SubwayStationDaegu    -2.509e+00  1.443e+00 -1.739  0.0820 .
YrSold:SubwayStationKyungbuk_uni_hospital -7.950e+00  7.476e-01 -10.634 < 2e-16 ***
YrSold:SubwayStationMyung-duk  -2.938e+00  7.337e-01 -4.004 6.30e-05 ***
YrSold:SubwayStationno_subway_nearby  1.937e+00  9.025e-01   2.147  0.0319 *
YrSold:SubwayStationSin-nam  -3.952e+00  9.119e-01 -4.334 1.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.16 on 5869 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8623
F-statistic: 1758 on 21 and 5869 DF,  p-value: < 2.2e-16

> AIC(regression1a)
[1] 61101.43
```

Exhibit 11 [Linear Regression With 6 Variables]

```
Call:
lm(formula = SqrtPrice ~ ApartmentSize * HallwayType + ApartmentSizeSquared +
  YrSold * SubwayStation + Amenities + TotalFacilities, data = CleanApartmentData)

Residuals:
    Min       1Q   Median       3Q      Max
-299.007  -22.658    2.106   23.314  152.073

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.179e+04  1.013e+03  -41.238 < 2e-16 ***
ApartmentSize    3.537e-01  7.470e-03   47.346 < 2e-16 ***
HallwayTypemixed -3.578e+01  5.878e+00  -6.088 1.22e-09 ***
HallwayTypeterraced -6.111e+01  6.046e+00 -10.107 < 2e-16 ***
ApartmentSizeSquared -1.111e-04  3.076e-06 -36.135 < 2e-16 ***
YrSold          2.084e+01  5.037e-01   41.375 < 2e-16 ***
SubwayStationBanwoldang 3.094e+04  1.360e+03   22.747 < 2e-16 ***
SubwayStationChil-sung-market 1.363e+04  2.745e+03    4.967 6.98e-07 ***
SubwayStationDaegu 6.273e+03  2.360e+03    2.658 0.00787 **
SubwayStationKyungbuk_uni_hospital 1.598e+04  1.222e+03   13.079 < 2e-16 ***
SubwayStationMyung-duk 8.595e+03  1.200e+03    7.162 8.92e-13 ***
SubwayStationno_subway_nearby 1.191e+04  1.504e+03    7.924 2.73e-15 ***
SubwayStationSin-nam 8.662e+03  1.490e+03    5.812 6.50e-09 ***
Amenities        1.658e+01  3.012e-01   55.051 < 2e-16 ***
TotalFacilities  -4.650e+00  2.819e-01 -16.497 < 2e-16 ***
ApartmentSize:HallwayTypemixed 5.525e-02  8.538e-03    6.471 1.05e-10 ***
ApartmentSize:HallwayTypeterraced 1.028e-01  8.254e-03   12.454 < 2e-16 ***
YrSold:SubwayStationBanwoldang -1.536e+01  6.760e-01 -22.717 < 2e-16 ***
YrSold:SubwayStationChil-sung-market -6.791e+00  1.364e+00 -4.977 6.63e-07 ***
YrSold:SubwayStationDaegu -3.121e+00  1.172e+00 -2.662 0.00779 **
YrSold:SubwayStationKyungbuk_uni_hospital -7.952e+00  6.073e-01 -13.094 < 2e-16 ***
YrSold:SubwayStationMyung-duk -4.264e+00  5.964e-01 -7.150 9.76e-13 ***
YrSold:SubwayStationno_subway_nearby -5.919e+00  7.470e-01 -7.923 2.75e-15 ***
YrSold:SubwayStationSin-nam -4.314e+00  7.406e-01 -5.824 6.05e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.06 on 5867 degrees of freedom
Multiple R-squared:  0.9095,    Adjusted R-squared:  0.9092
F-statistic: 2565 on 23 and 5867 DF, p-value: < 2.2e-16

> AIC(regression1a)
[1] 58652.14
```

Exhibit 12 [Floors vs Sqrt Floors Histogram and Scatterplot]

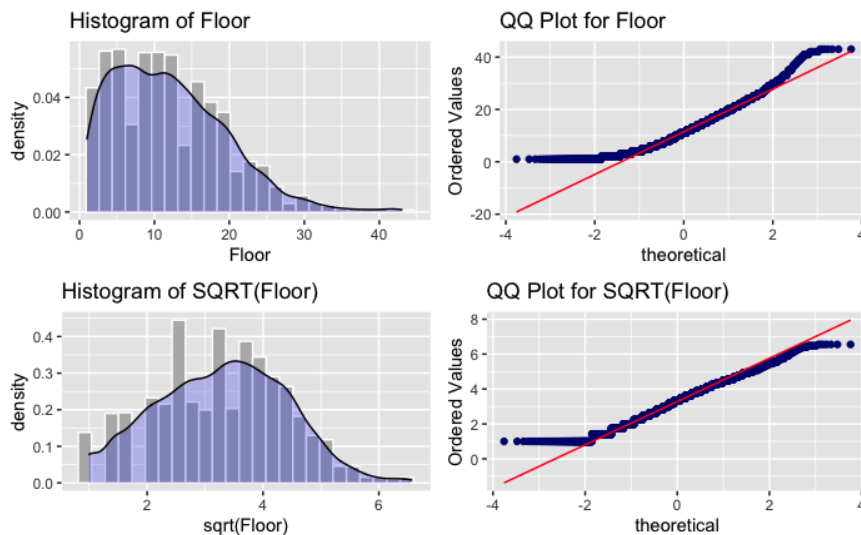


Exhibit 13 [Box Cox of Floors Post-Square-Root Transformation]

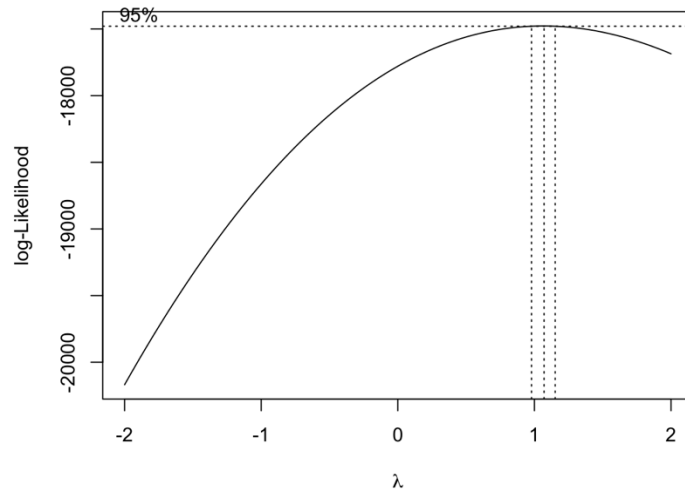


Exhibit 14 [Final Regression]

```
Call:
lm(formula = SqrtPrice ~ ApartmentSize * HallwayType + ApartmentSizeSquared +
    YrSold * SubwayStation + Amenities + TotalFacilities + SqrtFloors,
    data = CleanApartmentData)
```

Residuals:

Min	1Q	Median	3Q	Max
-285.257	-22.529	1.901	22.960	133.145

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.181e+04	9.858e+02	-42.413	< 2e-16 ***
ApartmentSize	3.596e-01	7.273e-03	49.442	< 2e-16 ***
HallwayTypemixed	-3.214e+01	5.721e+00	-5.617	2.03e-08 ***
HallwayTypeterraced	-5.297e+01	5.898e+00	-8.982	< 2e-16 ***
ApartmentSizeSquared	-1.104e-04	2.992e-06	-36.900	< 2e-16 ***
YrSold	2.084e+01	4.899e-01	42.534	< 2e-16 ***
SubwayStationBanwoldang	3.109e+04	1.323e+03	23.493	< 2e-16 ***
SubwayStationChil-sung-market	1.282e+04	2.670e+03	4.802	1.61e-06 ***
SubwayStationDaegu	6.806e+03	2.295e+03	2.965	0.00304 **
SubwayStationKyungbuk_uni_hospital	1.602e+04	1.189e+03	13.480	< 2e-16 ***
SubwayStationMyung-duk	8.730e+03	1.167e+03	7.479	8.57e-14 ***
SubwayStationno_subway_nearby	1.272e+04	1.463e+03	8.691	< 2e-16 ***
SubwayStationSin-nam	8.848e+03	1.450e+03	6.104	1.10e-09 ***
Amenities	1.539e+01	3.002e-01	51.248	< 2e-16 ***
TotalFacilities	-4.640e+00	2.742e-01	-16.922	< 2e-16 ***
SqrtFloors	7.672e+00	4.190e-01	18.312	< 2e-16 ***
ApartmentSize:HallwayTypemixed	4.575e-02	8.321e-03	5.498	3.99e-08 ***
ApartmentSize:HallwayTypeterraced	9.226e-02	8.049e-03	11.461	< 2e-16 ***
YrSold:SubwayStationBanwoldang	-1.543e+01	6.576e-01	-23.466	< 2e-16 ***
YrSold:SubwayStationChil-sung-market	-6.387e+00	1.327e+00	-4.812	1.53e-06 ***
YrSold:SubwayStationDaegu	-3.387e+00	1.140e+00	-2.970	0.00299 **
YrSold:SubwayStationKyungbuk_uni_hospital	-7.973e+00	5.907e-01	-13.498	< 2e-16 ***
YrSold:SubwayStationMyung-duk	-4.333e+00	5.801e-01	-7.468	9.31e-14 ***
YrSold:SubwayStationno_subway_nearby	-6.319e+00	7.269e-01	-8.692	< 2e-16 ***
YrSold:SubwayStationSin-nam	-4.408e+00	7.204e-01	-6.119	1.00e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.1 on 5866 degrees of freedom
Multiple R-squared: 0.9144, Adjusted R-squared: 0.9141
F-statistic: 2612 on 24 and 5866 DF, p-value: < 2.2e-16

```
> AIC(regression1a)
[1] 58326.67
```

Exhibit 15 [Project Workload Attachment]

	Aaron Yuan	Eddie Zhang	Tong Yu	Lisiman Hua	Yujia Wang
Business Understanding	X	X	X	X	X
Data Understanding	X	X	X		X
Data Preparation	X	X			X
Modeling	X	X	X	X	X
Evaluation	X	X			X
Deployment			X	X	
Risks & Considerations	X	X	X	X	
R Script/Code	X	X	X	X	X