STAT 151A Project

Predicting Housing Resale Prices in Singapore

Michelle Vuong, Celina Mac, Lewis Chong

April 10th 2024

Introduction

In many developed nations across the world, the real estate sector is an area of valuable business prospects due to the high demand of good land. Within the real estate sector, residential homes are a commodity extremely sought after and utilized, but are also incredibly big investments. In the lens of a real estate investor and prospective homeowner, such an investment requires a myriad of factors to consider in order to determine the best returns. With this in mind, it would be worthwhile to explore statistical approaches to assist with it. In this project, we aim to understand the Singaporean housing market and its trends (if any) from the years 2017-2022, which coincidentally involves the effects of the COVID-19 pandemic

Problem Description

For the purpose of this report, we will attempt to formulate an effective prediction model for resale prices in the country of Singapore using the tools associated with linear modelling which will mainly comprise of linear regression techniques. We will analyze how generally, housing prices may have shifted based on some features on the state of housing (housing type, size, and town) and whether or not they had a significant impact on resale prices. While we will keep in mind as researchers that the global pandemic had occurred within the time frame we are controlling, we will not be directly referencing COVID-19 as the reason for any of our findings and instead, attempt to account for that by taking into account inflation that may have occurred during that time. Our leading question for this project will now be: Was there any change in the housing market for Singaporean homes in the years between 2017-2022 in the selected towns and their respective most popular home types and rooms?

Data

Our data collection process began with an open web research on the housing markets of Singapore. We landed on Kaggle, an open source hub of public data sets uploaded by public users, which can be used for data exploration, building predictive models, and general practice with real-world data. Specifically, our data from Kaggle was transcribed by a user from the Singapore Government Agency Website that studied Resale flat prices based on registration date from Jan-2017 onwards. Data was collected by the Housing and Development Board, commonly referred to as "HDB". It is a statutory board of the Ministry of National Development in Singapore and it seeks to provide support in homeownership and ease in rental processes for residents. Simultaneously as the HDB are providing aid, they are collecting data on what home are being bought, built, and sold for. As this government established board provides public housing for more than 80% of Singapore's population, this project will make the assumption that all data was collected as a random sample of Singapore's population and data quality is up to par with research standards.

1. Exploratory Data Analysis and Data Preprocessing

We will begin with Exploratory Data Analysis to understand the data that we are working with.

```
# Data from the Housing and Development Board of Singapore
housing <- read_csv("Resale_Price_2017_2022.csv",show_col_types=FALSE)
```

Data Inspection

```
## Datasets contains NA values :  FALSE
```

We have found that there are no null values in the dataset, and so it is clean for further analyses that will be conducted for this report.

```
# Inspection using the glimpse() function
glimpse(housing)
```

```
## Rows: 134,168
## Columns: 11
## $ month               <chr> "2017-01", "2017-01", "2017-01", "2017-01", "2017-~
## $ town                <chr> "ANG MO KIO", "ANG MO KIO", "ANG MO KIO", "ANG MO ~
## $ flat_type           <chr> "2 ROOM", "3 ROOM", "3 ROOM", "3 ROOM", "3 ROOM", ~
## $ block               <chr> "406", "108", "602", "465", "601", "150", "447", "~
## $ street_name         <chr> "ANG MO KIO AVE 10", "ANG MO KIO AVE 4", "ANG MO K~
## $ storey_range        <chr> "10 TO 12", "01 TO 03", "01 TO 03", "04 TO 06", "0~
## $ floor_area_sqm      <dbl> 44, 67, 67, 68, 67, 68, 68, 67, 68, 67, 68, 67, 67~
## $ flat_model          <chr> "Improved", "New Generation", "New Generation", "N~
## $ lease_commence_date <dbl> 1979, 1978, 1980, 1980, 1980, 1981, 1979, 1976, 19~
## $ remaining_lease     <chr> "61 years 04 months", "60 years 07 months", "62 ye~
## $ resale_price        <dbl> 232000, 250000, 262000, 265000, 265000, 275000, 28~
```

Upon our first inspection of the data, we will be using the `resale_price` as the response variable since we are interested in predicting housing resale price. As for our potential regressors of interest, we are electing to utilize `month`, `town`, `flat_type`, `floor_area_sqm`, and `flat_model` variables and we will subset the data to only focus on these variables. These variables are defined as follows:

- `month`: The month at which the information on the particular home is recorded.
- `town`: The town this particular home is situated in.
- `flat_type`: The type of home described as either 1, 2, 3, 4, or 5 rooms, executive, or multi-generation.
- `floor_area_sqm`: The size of the home in regards to the area of the floor in squared meters.
- `flat_model`: The model of the home described as either 2-room, 3Gen, Adjoined flat, Apartment, DBSS, Improved, Improved-Maisonette, Maisonette, Model A, Model A-Maisonette, Model A2, Multi Generation, New Generation, Premium, Apartment Premium, Apartment Loft, Premium Maisonette, Simplified, Standard, Terrace, Type S1, Type S2.

The reason why we chose these variables in particular is because we believe they will be the most likely to have the most significant effect on housing resale price. This will be further explored as move on.

We also must note that for `town`, there are many distinct values:

```
cat("Distinct values for `town` :",length(unique(housing$town)))
```

```
## Distinct values for ‘town‘ : 26
```

This is important to note as it may affect the effectiveness of our model which will aim to account for soon in this report.

2. Data Manipulation

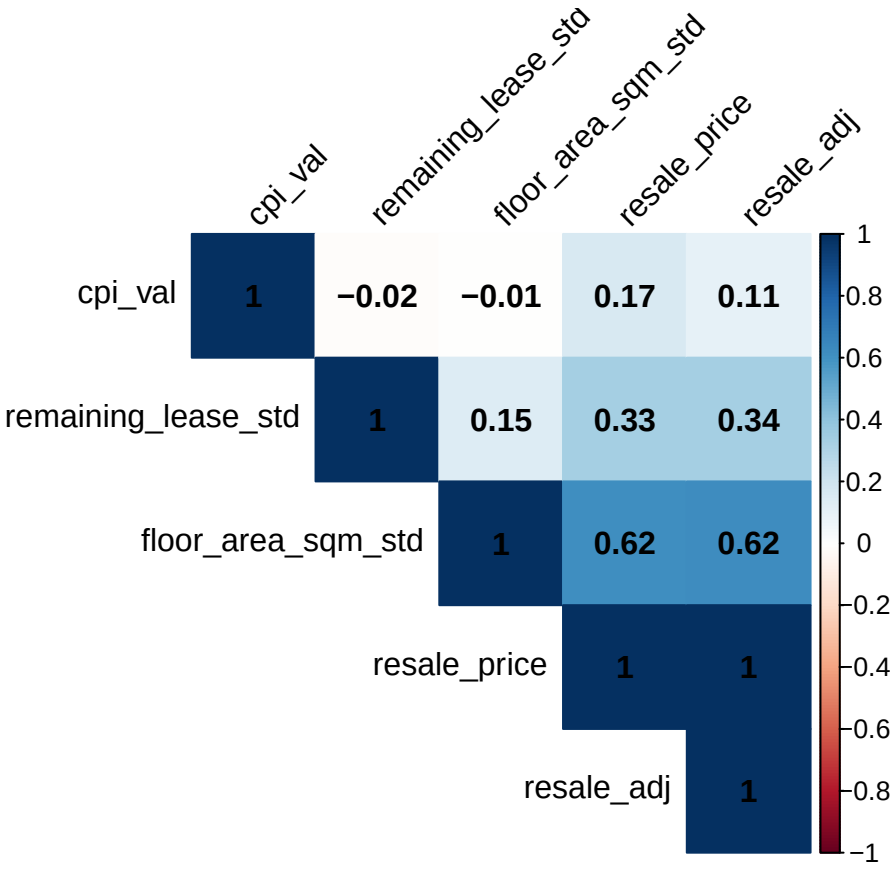
To better ease the formulation of our model, we decided to convert the `remaining_lease` column that contains how long the lease is to be of unit months instead of the current year+month.

Next we would like to make adjustments for inflation using the Singaporean Consumer Price Index for "Housing and Utilities". This main purpose for this adjustment is to account for the time variable since we recognize that COVID-19 could possibly have an effect on housing prices. We will be calculating the "real," or "relative" price of housing in Singapore, which in the field of Economics, is used to adjust for inflation. We will use data on the Consumer Price Index in Singapore provided by the Department of Statistics in Singapore.

3. Standardizing Predictor Columns.

We noticed from inspecting our data that the variable `floor_area_sqm` which describes how big the house is in terms of the area in square meters, contains extremely big values compared to everything else. Additionally, after conversions, the remaining lease in months corresponding to each house is also quite large. Our concern is that these large values may dominate other variables that we may want to use for our model.Thus, we elected to standardize these predictor columns in preparation for use later on in our model.

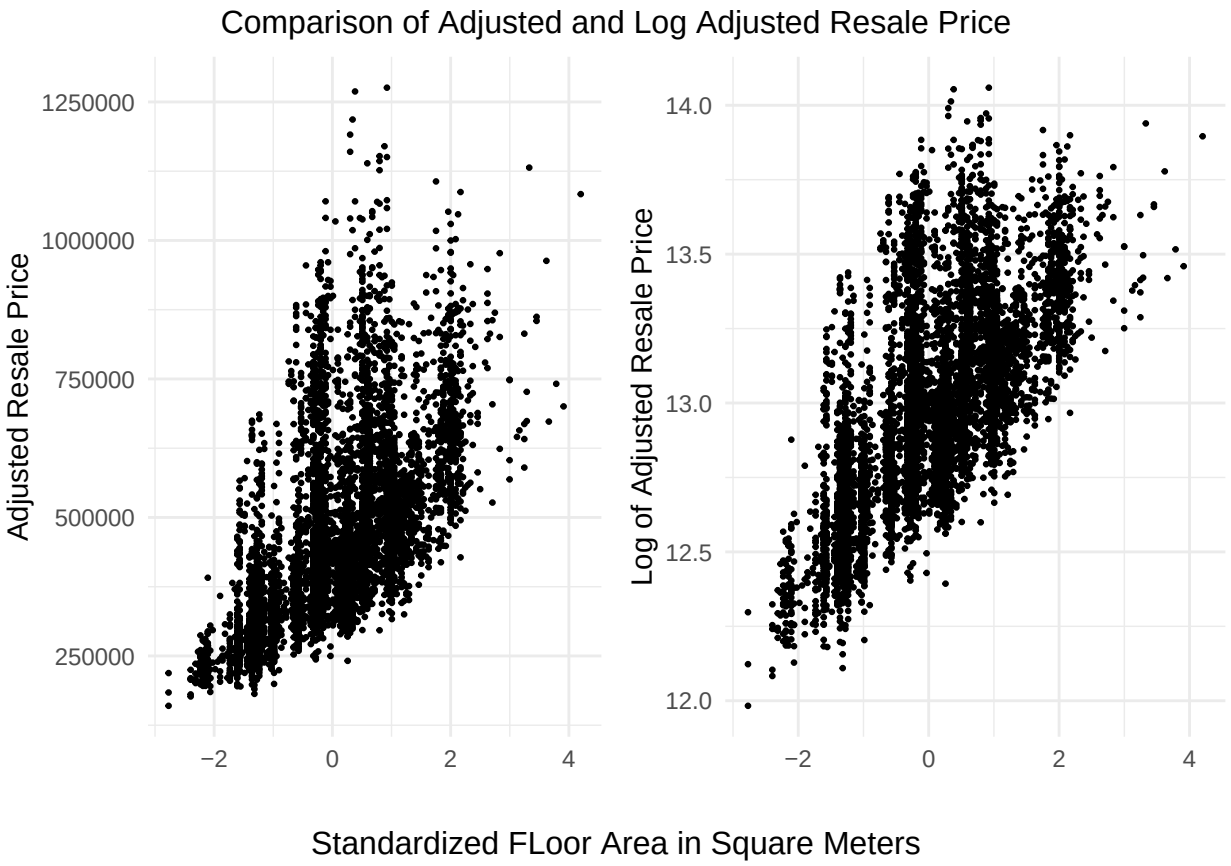
And now, we will check the correlations of these variables with each other to see if there any indications of multicolinearity, so that we can be sure to account for later on when formulating our prediction model.



Upon analyzing the correlation plot above and further considering all the data manipulation and analysis we have done up until this point, we decided that we will look into the variable `floor_area_sqm_std` as our regressor for our model. This is because `resale_adj` and `floor_area_sqm_std` have the highest correlation with each other compared to the other variables of interest.

4. Log Transforming our Resale Data and their Plots

Here, we beginning plotting our resale data with the `floor_area_sqm_std` variable since we want to see how resale data is associated with the size of the home. Additionally, we elected to log transform our resale data and plotted it against the area of the home in square meters in order to reduce any possible skewness of the data. Below is the two plots side-by-side for comparison. Note that these plots were conducted by first taking a random sample of 7000 homes from the dataset to prevent overcrowding and better visualize if there is any association between `resale_adj` and `floor_area_sqm_std`.



5. Methods

Now, we can begin formulating our prediction model. Based on previous attempts (that will be explained in the *Limitations* section of this report) we are choosing to select thirteen randomized towns and filter our data to the most popular housing models and house types for the selected towns in the country of Singapore. Since there are 26 towns in total and the amount of data we have is very large, we want to take a smaller sample of our data and perform regression analysis on it instead of utilizing the whole dataset. While having more data is better than having too little, we realize that prediction models for our scenario is very complicated and regression will likely not be able to capture many of the intricacies of the housing market.

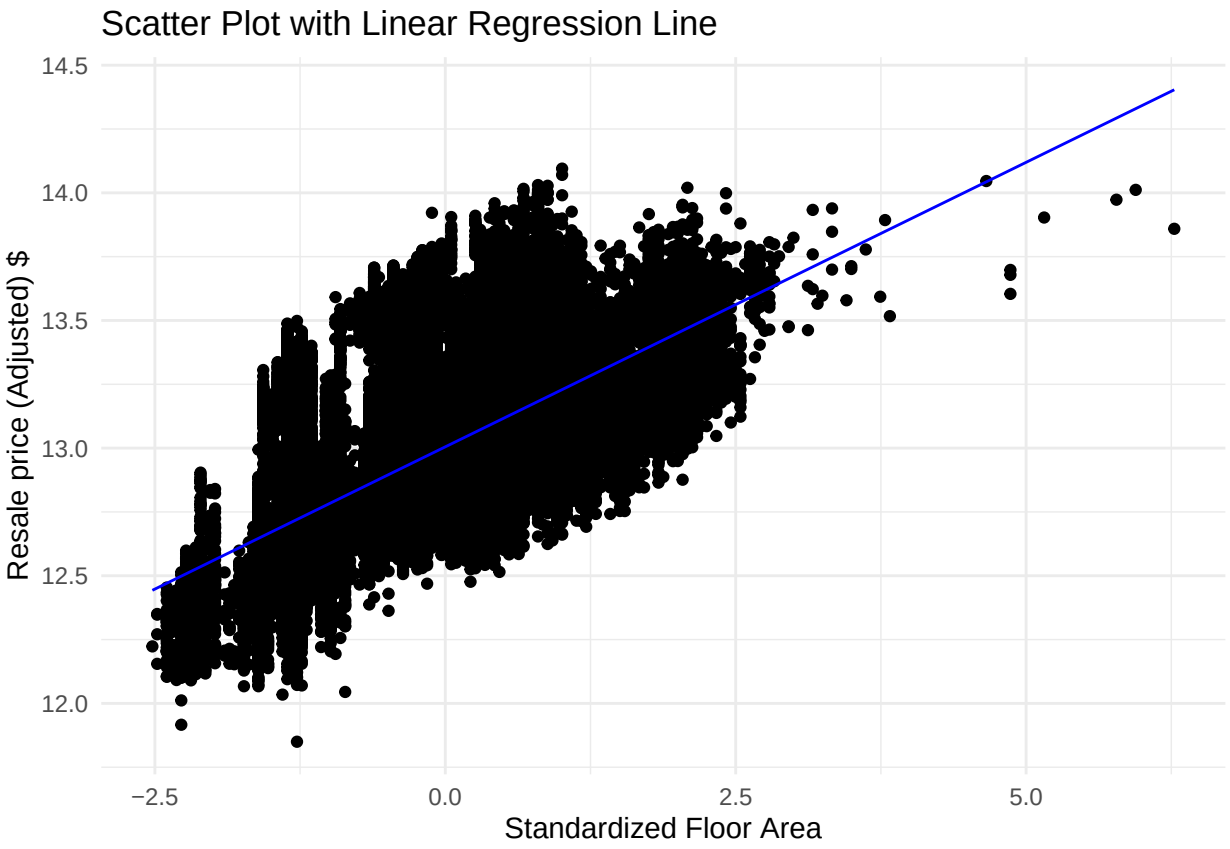
```
# Selecting random sample of towns in Singapore from the dataset
set.seed(248)

randomTown <- sample(unique(housing$town), 13, replace=FALSE)

# Filtering the data based on the specific towns used
housing_dat <- housing %>%
  dplyr::filter(town %in% randomTown)

# Finding the most popular model and type of home for these towns
top_model <- names(sort(table(housing_dat$flat_model), decreasing = TRUE)[1])
top_type <- names(sort(table(housing_dat$flat_type), decreasing = TRUE)[1])
```

Of the thirteen towns, the top housing model is “Model A” and a house with “4 Rooms”. Now that we know know the most populous buildings in the most populous towns, we will run our regression on the log

adjusted resale prices that account for inflation.

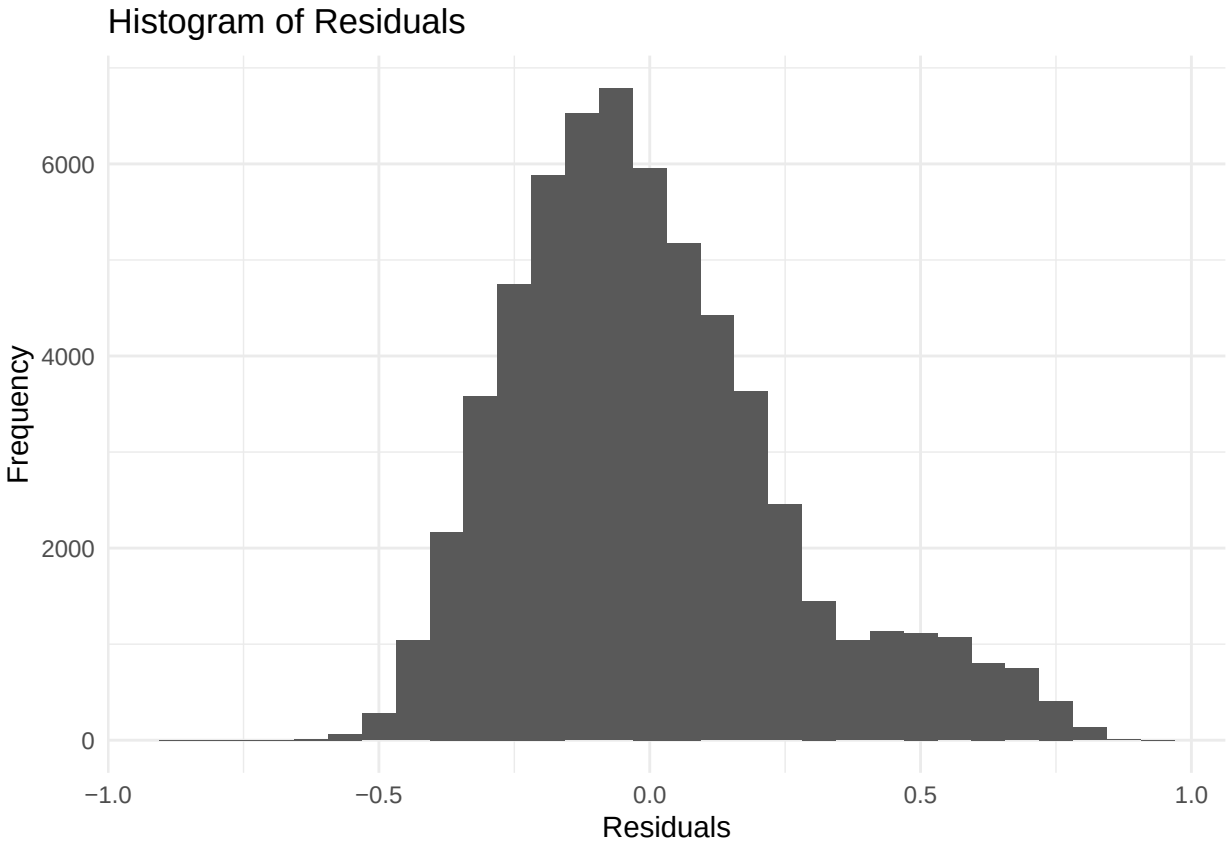
```
##
## Call:
## lm(formula = log(resale_adj) ~ floor_area_sqm_std, data = housing_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87056 -0.18339 -0.03889  0.13898  0.94240
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        13.004994   0.001049 12393.3   <2e-16 ***
## floor_area_sqm_std  0.222880   0.001043   213.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2582 on 60631 degrees of freedom
## Multiple R-squared:  0.4298, Adjusted R-squared:  0.4298
## F-statistic: 4.57e+04 on 1 and 60631 DF,  p-value: < 2.2e-16
```

The summary statistics of our model give us an R^2 of 0.4562, and a model that claims a one unit increase in floor area square meter is predicted to result in the log adjusted resale price to increase by 0.244. For this scenario, we believe this interpretation of the coefficient for `floor_area_sqm_std` is valid to an extent as intuitively, a larger home would generally result in a higher price as it occupies more land. Below is a visualization of our model.

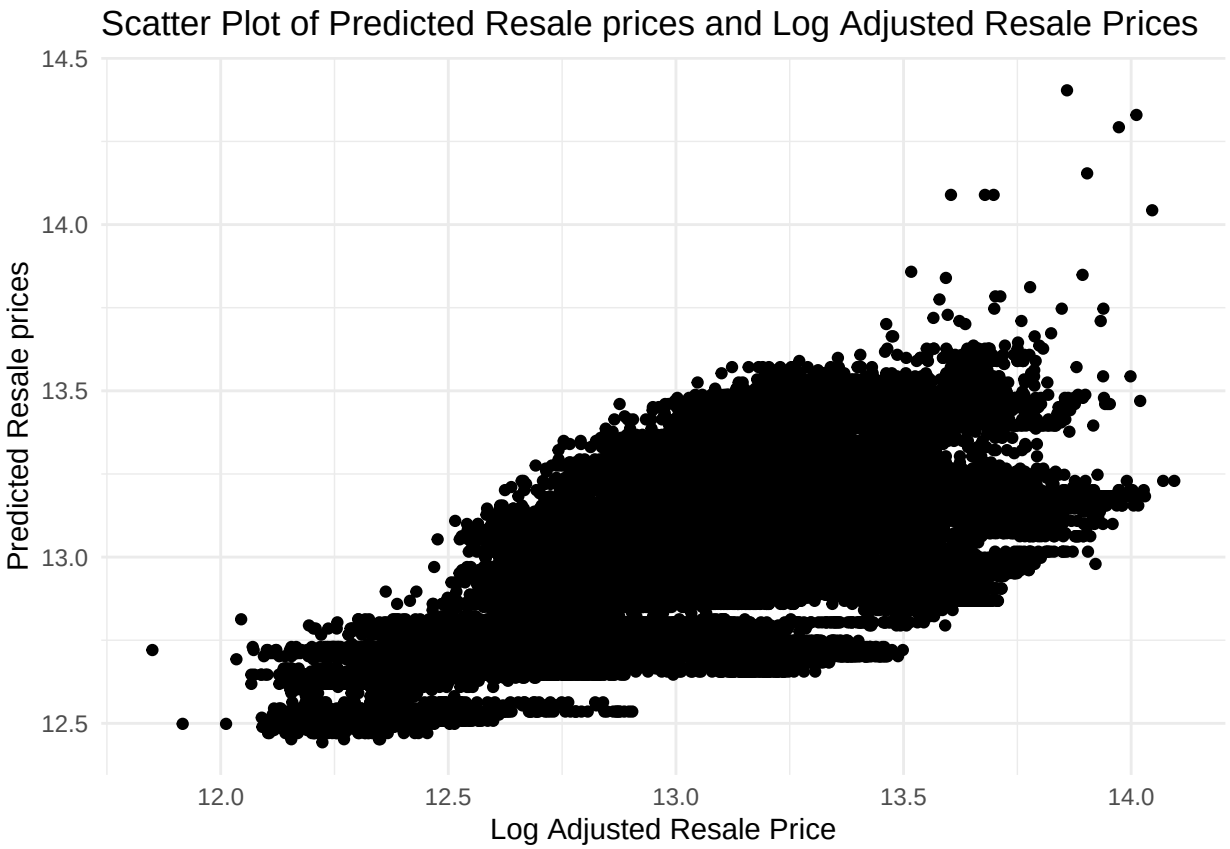


Below is a histogram of the residuals from our regression model. From the looks of it, the residuals are roughly normally distributed and indicates that any normality assumption is likely to be true.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



And thus, we may proceed with the comparison of log resale prices with adjusted inflation and predicted values.



Results and Discussion

Based on our linear regression (lm()) model results and the plots above, it is somewhat reasonable to claim that the logged resale adjusted-for-inflation prices follow a linear trend. There are clearly horizontal trends and intense clustering in the plot right above, however, these problems will be addressed in the limitations of the extent of our project below. However, narrowing down to the scatter plot of floor area per square meter and adjusted resale prices, there is a slight increasing linear trend in the plot with large amounts of concentration between 0-2.5 square meters. This means that as the size of the home increases, so does the log of the resale prices. This makes sense intuitively due to the nature of homes, the larger they are, the more land they take up and the more you have to pay for the land.From our linear model summary with floor square meter as the only predictor, it produced to us a R squared value of 0.4298 which indicates it is not the best predictor, but after running f-tests on each of the other predictors given in our data set, this came out to be the most significant. When we checked our residuals given from our model, we noticed that the histogram was very close to being normally distributed which means that we achieved homoscedasticity and had a decent linear model with floor square meter as the predictor. This normality assumption also allowed us to make inference comments on the interpretation of our model.

Limitations and Future Work

The linear regression model we have above is unlikely to be considered a good prediction model for housing prices for many reasons. For one, we only utilized the size of the home in terms of the floor area in squared meters to formulate our prediction model using linear regression. As explained above, it makes sense intuitively why a larger home would generally result in a higher resale price compared to smaller homes. Additionally, we also controlled the other variables that would affect housing prices in our model

which were the type of the flat and the model of the flat. In the end, we only really ended up with a model that depicts the relationship between the size of the home and the resale price of the home of a particular flat type and flat model, which in our case are 4 room Model A flats.

The housing market is very complex and the factors that contribute to determining the sale price of homes includes more than just the size of the home which we unfortunately were unable to capture in our linear regression model with the tools we were equipped with at the extent of our knowledge on linear modelling.

We should note that an attempt had been made prior to the procedure and results we have gone through at the top in regards to formulating a prediction model that takes into account more than just the size of the home as predictors for housing resale price. We will now explain here:

Below is all work prior to our consultation (Thursday, 5/2/2024) and new formulated question. We did not want to delete all of our work prior to the consultation because our original question had been approved many weeks prior to the meeting before being advised that the question was not answerable with our given data set. However, we did learn many things about our categorical variables from this process and want to show them below.

===

We have noted for this data set that it contains a lot of categorical variables and the variables that we have chosen also fall under this characterization and so we must employ a technique that will be able to translate our categorical variables into numerical data for the purpose of our models.

1. One-hot-encoding From our data inspection, we see that `flat_type` is a categorical variable. Thus, we will applied one hot encoding on it for the model to regress.

```
# housing <- housing %>%
#   mutate(
#     is_2_room = ifelse(flat_type == "2 ROOM", 1, 0),
#     is_3_room = ifelse(flat_type == "3 ROOM", 1, 0),
#     is_4_room = ifelse(flat_type == "4 ROOM", 1, 0),
#     is_5_room = ifelse(flat_type == "5 ROOM", 1, 0),
#     is_executive = ifelse(flat_type == "EXECUTIVE", 1, 0),
#     is_1_room = ifelse(flat_type == "1 ROOM", 1, 0),
#     is_multi_generation = ifelse(flat_type == "MULTI-GENERATION", 1, 0)
#   ) %>%
#   dplyr::select(!flat_type)
```

As noted near the beginning, we see that there are a lot of distinct values for our `town` variable which may affect our prediction model.

In order to remedy this in a way that will be better utilized for our model, we categorized the different towns of Singapore into NSEW (North, South, East, West) regions, and further applied one-hot encoding as done above since it is a categorical variable.

```
# Function to categorize towns into NSEW regions
# categorize_town <- function(town) {
#   north <- c("ANG MO KIO", "SEMBAWANG", "SENGKANG", "WOODLANDS", "YISHUN","BISHAN")
#   south <- c("BUKIT MERAH", "BUKIT TIMAH", "CENTRAL AREA", "QUEENSTOWN")
#   east <- c("BEDOK", "MARINE PARADE", "PASIR RIS", "TAMPINES")
#   west <- c("BUKIT BATOK", "BUKIT PANJANG", "CHOA CHU KANG", "CLEMENTI", "JURONG EAST", "JURONG WEST"
#
#   if (town %in% north) {
#     return("North")
#   } else if (town %in% south) {
```

```
#     return("South")
#   } else if (town %in% east) {
#     return("East")
#   } else if (town %in% west) {
#     return("West")
#   } else {
#     return("Other")
#   }
# }
#
# # Add a new column for NSEW region
# housing <- housing %>%
#   rowwise() %>%
#   mutate(region = categorize_town(toupper(town))) %>%
#   mutate(
#     is_north = ifelse(region == "North", 1, 0),
#     is_south = ifelse(region == "South", 1, 0),
#     is_west = ifelse(region == "West", 1, 0),
#     is_east = ifelse(region == "East", 1, 0)
#   ) %>%
#   dplyr::select(!region)
```

We noted that for one-hot encoding on the columns `region` and `flat_type` , we can just drop one of the columns as it can be identified by the rest of the columns, i.e `(is_north,is_south,is_west) = (0,0,0)`corresponds to the house being in the East Region.

```
# housing <- housing %>%
#         dplyr::select(!c(is_east,is_multi_generation))
```

In order to evaluate the most statistically significant housing metrics and reduce the number of predictors we have, we used regularization. We began with Ridge Regularization as shown below:

```
# df <- housing %>% dplyr::select(!month)
#
# ori <- lm(log(resale_adj) ~ . + 0, data = df)
#
# yhat <- predict(ori)
#
# y <- log(df$resale_adj)
#
# MSE_ori <- mean((y-yhat)^2)
# cat("MSE from the linear model: ", MSE_ori)
#
# epsilon <- data.frame(resid=resid(ori),month=housing$month)
#
# residuals_means <- epsilon %>%
#   group_by(month) %>%
#   summarise(mean_residual = mean(resid))
#
# # Now you can plot the mean residuals against the months
# ggplot(residuals_means, aes(x = month, y = mean_residual)) +
#   geom_line() +
#   labs(x = "Month", y = "Mean Residuals", title = "Mean Residuals Over Time")
```

Ridge and Lasso

```
# # Set seed for reproducibility
# set.seed(123)
# # Create an index for splitting the data
# split <- createDataPartition(y = housing$resale_price, p = 0.8, list = FALSE)
# X <- dplyr::select(df,!resale_price)
# X_train <- X[split,]
# X_test <- as.matrix(X[-split,])
# y_train <- df$resale_price[split]
# y_test <- df$resale_price[-split]
# # Fit Ridge model
# ridge <- glmnet::glmnet(x = X_train, y = y_train, alpha = 0)
# yhat_ridge <- predict(ridge, newx = X_test)
# lambda_ridge <- ridge$lambda
# # Min MAE
# mae_ridge <- apply(abs(yhat_ridge - y_test),2,mean)
#
# #col mean for each lambda
# min_ridge <- which.min(mae_ridge) #index corresponding to min MAE
# min_lambda_ridge <- lambda_ridge[min_ridge]
# coef_min_ridge <- coef(ridge)[,min_ridge] #Corresponding indices of the non-zero coefs
# lagged_features <- as.numeric(which(coef_min_ridge != 0))
# print("Lagged features present in the MAE-optimal ridge model:")
#
# # Fit lasso model
# lasso <- glmnet::glmnet(x = X_train, y = y_train, alpha = 1)
# yhat_lasso <- predict(lasso, newx = X_test)
# lambda_lasso <- lasso$lambda # Min MAE
# mae_lasso <- apply(abs(yhat_lasso - y_test),2,mean)
# #col mean for each lambda
# min_lasso <- which.min(mae_lasso) #index corresponding to min MAE
# min_lambda_lasso <- lambda_lasso[min_lasso]
# coef_min_lasso <- as.data.frame(coef(lasso)[,min_lasso])
# #Corresponding indices of the non-zero coefs
# coef_min_lasso
```

F-test for feature selection

Using the F-test to understand the variability between the features in the data set, we want to select the features with most impact on our y variable.

```
# ANOVA - This is giving high RSE, meaning there is a lot of unexplained variability.
# anova_test <- aov(resale_price ~ is_1_room +
#                     is_2_room +
#                     is_3_room +
#                     is_4_room +
#                     is_5_room +
#                     is_executive +
#                     is_north +
#                     is_south +
#                     is_west +
#                     floor_area_sqm_std +
```

```
#                     remaining_lease_std, df)
#
#
# # F-test on each features with 2 levels
# is_1_room <- var.test(resale_price ~ is_1_room, df,
#          alternative = "two.sided")
# is_2_room <- var.test(resale_price ~ is_2_room, df,
#          alternative = "two.sided")
# is_3_room <- var.test(resale_price ~ is_3_room, df,
#          alternative = "two.sided")
# is_4_room <- var.test(resale_price ~ is_4_room, df,
#          alternative = "two.sided")
# is_5_room <- var.test(resale_price ~ is_5_room, df,
#          alternative = "two.sided")
# is_executive <- var.test(resale_price ~ is_executive, df,
#          alternative = "two.sided")
# is_north <- var.test(resale_price ~ is_north, df,
#          alternative = "two.sided")
# is_south <- var.test(resale_price ~ is_south, df,
#          alternative = "two.sided")
# is_west <- var.test(resale_price ~ is_west, df,
#          alternative = "two.sided")
#
# F_test_vals <- list(
#   is_1_room = is_1_room,
#   is_2_room = is_2_room,
#   is_3_room = is_3_room,
#   is_4_room = is_4_room,
#   is_5_room = is_5_room,
#   is_executive = is_executive,
#   is_north = is_north,
#   is_south = is_south,
#   is_west = is_west
# )
#
# # Extract p-values
# p_vals <- sapply(F_test_vals, function(result) {
#   result$p.value
# })
#
# # Check if any p-value is less than a significance level (e.g., 0.05)
# any(p_vals > 0.05)
# big_p <- which(p_vals>0.05)
#
# cat("Feature with p_value > 0.05: ", names(df)[big_p],"\n")
# cat("Corresponding p-value: ",p_vals[big_p])

# More than two levels, won't run with F-test.
#floor_area_sqm_std <- var.test(resale_price ~ floor_area_sqm_std, housing,
     #    alternative = "two.sided")
#remaining_lease_std <- var.test(resale_price ~ remaining_lease_std, housing,
     #    alternative = "two.sided")
```

In the Lasso model the feature `is_5_room` has a zero coefficient and this F-test confirms that we can drop this feature due to the p-value being higher than 0.05. So we will re-run the linear regression ommitting this feature.

```
# ori_no_5_rooms <- lm(log(resale_price) ~ is_2_room +
#                                          is_3_room +
#                                          is_4_room +
#                                          is_executive +
#                                          is_1_room +
#                                          is_north +
#                                          is_south +
#                                          is_west +
#                                          floor_area_sqm_std +
#                                          remaining_lease_std +
#                                          0, data = df)
#
# rsquared <- summary(ori_no_5_rooms)
# rsquared$r.squared
```

Regression Analysis on the Effects of COVID-19 on the Housing Market in Singapore

Now that we have performed ANOVA and F-test for feature selection and determined which variables to utilize in our model, we can move forward with further analyzing the effects of COVID-19 on the housing market in Singapore.

Subsetting our dataset

We will begin by filtering all dates before April 2020 as that was when the Singaporean government began enforcing preventive measures for the pandemic. This date was chosen based on the following information:

1. The first COVID-19 case in Singapore was confirmed on January 23, 2020.
2. COVID-19 clusters in the population were recorded in late March and early April 2020.
3. Singapore enacted the "COVID-19 Control Order" in April 3, 2020 and announced the "circuit breaker lockdown," which was a set of stringent preventive measures to curb the spread of COVID-19.

With all this information in mind, we believe that April 2020 would be the best date to choose as the boundary when subsetting our data into pre-COVID and COVID time periods.

```
# Subsetting our original dataset into the two time periods
pre_covid <- housing %>%
  filter(month < "2020-04-01")

covid <- housing %>%
  filter(month >= "2020-04-01")
```

Running Linear Regression on Both Periods

```
# Linear Regression on Pre-Covid Data
# pre_covid_lm <- lm(log(resale_price) ~ is_2_room +
#       is_3_room +
#       is_4_room +
#       is_executive +
#       is_1_room +
#       is_north +
#       is_south +
#       is_west +
#       floor_area_sqm_std +
#       remaining_lease_std +
#       0, data = pre_covid)
#
# # Linear Regression on Covid Period Data
# covid_lm <- lm(log(resale_price) ~ is_2_room +
#       is_3_room +
#       is_4_room +
#       is_executive +
#       is_1_room +
#       is_north +
#       is_south +
#       is_west +
#       floor_area_sqm_std +
#       remaining_lease_std +
#       0, data = covid)
#
# summary(pre_covid_lm)
# summary(covid_lm)
```

Results and Discussion

After sub-setting the data into two categories (pre-covid and post-covid) and running our regression on both categories, we called the R function summary() to gather information on how our model was performing. In our pre-covid model, we achieved a R squared value of 0.9484. In our post-covid model, we achieved a R squared of 0.9462. There is only a slight difference between the two, which tells us. . . .

Did not work because high R^2 value did not match the high residual error.

==========

There were multiple issues with this attempt:

1) Having such a high R^2 on housing prices is surprising, and is an indication of potential overfitting. As mentioned before, there are a lot of features of a house that affect its price but are not captured in our regressors. Since our R^2 was 0.99, that indicates that there are no such features and is perhaps the case in our attempt.

2) The size of the residuals in the plot do not look consistent with an R^2 of 0.99. Residuals should not have a marked time trend which indicates a violation of the normality assumption and was clearly present in our plot. We have learned that it may be due to our neglect for controlling for important variables and/or we are not using a valuable predictor for predicting housing resale prices.

References Resale flat prices based on registration date from Jan-2017 onwards. Data.gov.sg. https://beta.data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view

Origin of data set for Singaporean homes. Kaggle.com. https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression/notebook

Singaporean Consumer Price Index. https://www.singstat.gov.sg/whats-new/latest-news/cpi-highlights