STAT 151A Project

Predicting Housing Resale Prices in Singapore

Michelle Vuong, Celina Mac, Lewis Chong

April 10th 2024

Introduction

new test comment

Research Objectives

Data Collection

EDA + Data Preprocessing

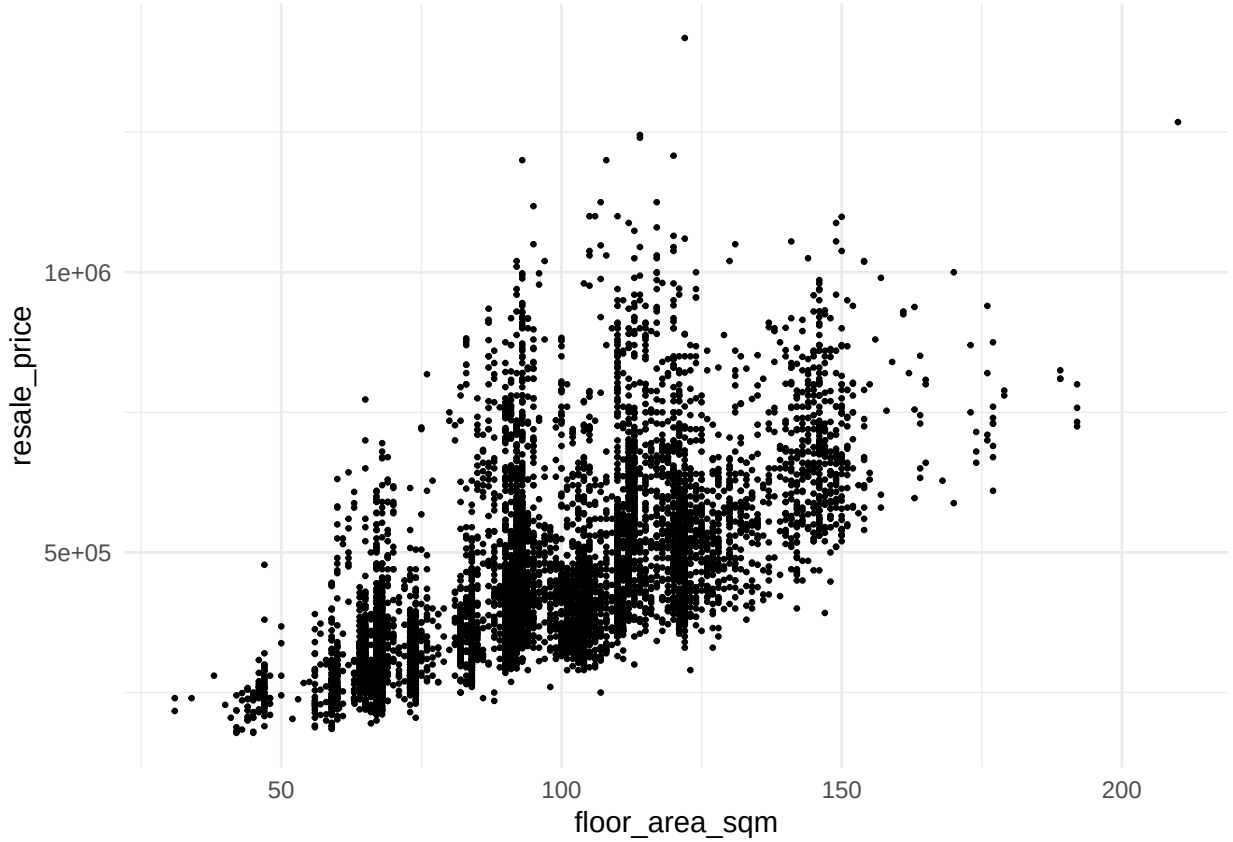
-using EDA to make decisions about the data : remove outliers, taking log of resale prices , removing certain categorical columns

1. Log transfrom

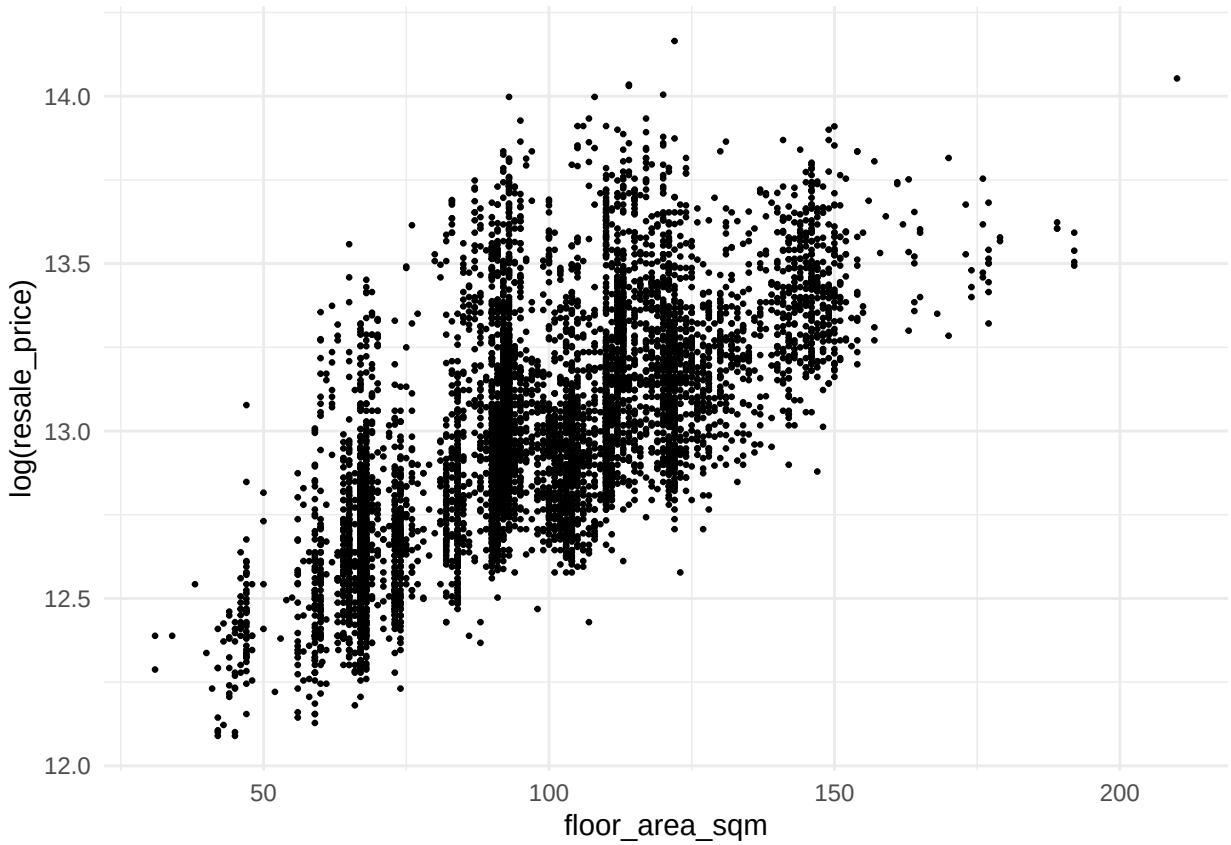
```
housing <- read.csv("Resale_Price_2017_2022.csv")

id <- sample(nrow(housing),7000)
sample_housing <- housing[id,]

## histogram
ggplot(sample_housing) +
  geom_point(aes(x=floor_area_sqm,y=resale_price),size=0.5,position="identity") +
  theme_minimal()
```




```
ggplot(sample_housing) +
  geom_point(aes(x=floor_area_sqm,y=log(resale_price)),size=0.5,position="identity") +
  theme_minimal()
```



We do a sample of 7000 on the original dataset, to argue that the increase of a small amount of floor area(sqm) doesn't result it a linear amount of resale price being added, but instead some non-linear increase in the price. This is equivalent to adding to a log of the resale prices. So we conclude that it results in better prediction if we do a regression on the log(resale price).

2. One hot encoding for Flat Type

We will be apply one hot encoding for the `flat_type` column to regress on the categorical values

```
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)
  )
```

###3. Data Manipulation We will be converting the `remaining_lease` column that contains how long the lease is to be of unit month instead of the current year+month.

```
##Function to convert from years+ months to months
extract_months <- function(duration_str) {
    # Split into components
    components <- strsplit(duration_str, " ", perl = TRUE)[[1]]

    # Extract years and months (if available)
```

```
    years <- as.numeric(components[1])
    months <- ifelse(length(components) >= 3, as.numeric(components[length(components)-1]), 0)

    # Return total months
    return(years * 12 + months)
}

housing <- housing %>%
            rowwise() %>%
            mutate(remaining_lease_mth = extract_months(remaining_lease))
```

Then, we will categorize the different towns of Singapore into NSEW regions:

```
# 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)))
```

Then, we apply one-hot encoding on the regions as well:

```
housing <- housing %>%
  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)
  )
```

Model Training and Evaluation

Limitation and Future Work