STAT 151A Project

Predicting Housing Resale Prices in Singapore

Michelle Vuong, Celina Mac, Lewis Chong

April 10th 2024

Introduction

Research Objectives

Data Collection

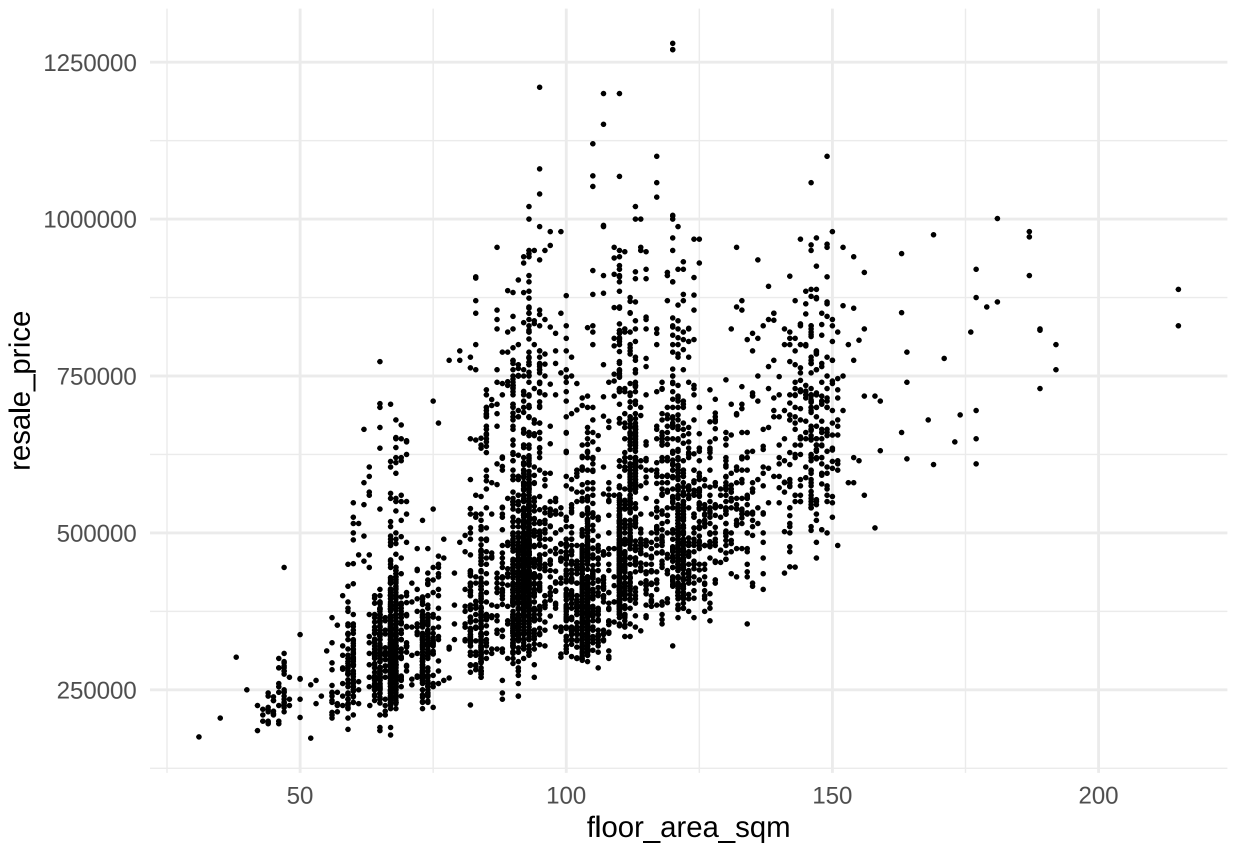
EDA

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

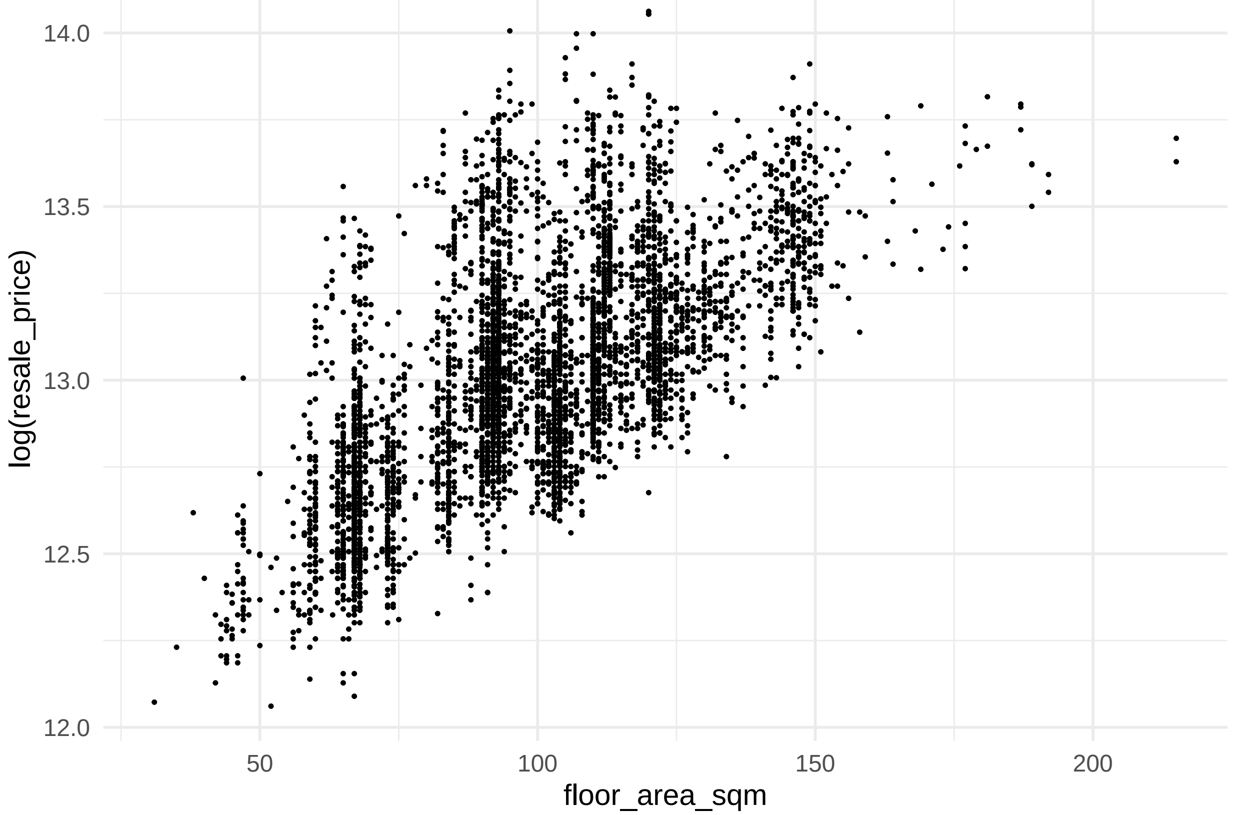
1. Log transfrom

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

id <- sample(nrow(housing),5000)
sample_housing <- housing[id,]
## histogram
ggplot(sample_housing) +
  geom_point(aes(x=floor_area_sqm,y=resale_price),size=0.3) +
  theme_minimal()
```



```
ggplot(sample_housing) +
  geom_point(aes(x=floor_area_sqm,y=log(resale_price)),size=0.3) +
  theme_minimal()
```



```
##Things to do
#drop some columns
#one hot encoding for flat type and flat model
# convert the leasing to months only
```

We do a sample of 5000 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).

Data Preprocessing

Model Training and Evaluation

Limitation and Future Work