final-project

Pragyat Agrawal Ruxuan Ji Meng-Chuan Chang

Abstract

In our final project, we would analyze rental data in the US, and build models to predict the rental price based on several factors, such as the number of dish-washer or the population.

Data Preprocessing

Read the data

Firstly, we read the data from Kaggle - US Rental Listings Summer 2021

```
data <- fread("data/Rental_Properties.csv")
summary(data)
```

Filter the data

The original dataset contains 276757 data, but we just need partial data. Before we randomly pick 20000 for further analysis, we can remove rows that is lack of important factors. The criteria is as follows

- sqft (squart feet) must be non-zero
- population and the density must be non-zero
- price must be non-zero

```
data_filter <- data[(data$sqft!=0 & data$Population!=0),]
data_filter <-
  data_filter %>%
  drop_na(price)
set.seed(1)
data_20000 <- sample_n(data_filter, 20000)
```

Then we drop several columns which is clearly not helpful for predicting the rental price

- zip
- link
- street_address
- full_address
- ZipCity

```
data_20000_filter <-
  data_20000 %>%
  select(-zip, -link, -street_address, -full_address, -ZipCity)
```

Finally, we fill all NA with 0. The columns having NA is as follows

- pool
- dishwasher
- washer-dryer
- ac

- parking

Then we export the cleaned dataframe to csv

```r
data_20000_filter[is.na(data_20000_filter)] <- 0
summary(data_20000_filter)

file_path <- "data/Rental_Properties_20000.csv"

if(!file.exists(file_path)) {
  write.csv(data_20000_filter, file_path)
} else {
  data_20000_filter <- fread(file_path)
}
```

EDA

Model Training

Performance Analysis

Conclusion