# final-project

Pragyat Agrawal Ruxuan Ji Meng-Chuan Chang

## Contents



# 1 Executive Summary

## 1.1 Background

The US rental market has been growing rapidly growing over time, making it one of the most sought after areas for investment. Be it from small home owners to big private equity firms, everyone seems to be after housing, expecting the values of these houses to rise and supplement their income by renting these places. Many people consider that location is the "only important" factor responsible for a house's value and the rent that can be expected of it, but this is far from the truth. There are a lot of other factors that need to be considered for determining housing valuations as we see a huge variation in prices in houses located in the same vicinity. There must be something about these houses which is causing such a big price change. Hence we will be analyzing the data related to US Rental Listings in Summer of 2021, to find which of these factors, which consist of many in-house amenity components, impacts housing values the most.

These amenities range from simple aka micro-factors like the availability of Pools and Dishwashers in the house to major aka macro-factors i.e. cities. The data will also give us the opportunity to find in which cities are these factors playing the most impact. With over 27,000 values for each predictor in our data, we have a sufficient sample size to make a reasonable conclusion regarding the price of these summer rentals.

Graph: Rental vacancy rates in the United States from 2000 to 2021, by region (Source: Statista.com)

This shows how the US housing market has been more sought after year by year, making housing a form of valuable investment. This increase in demand has already pushed up the prices.

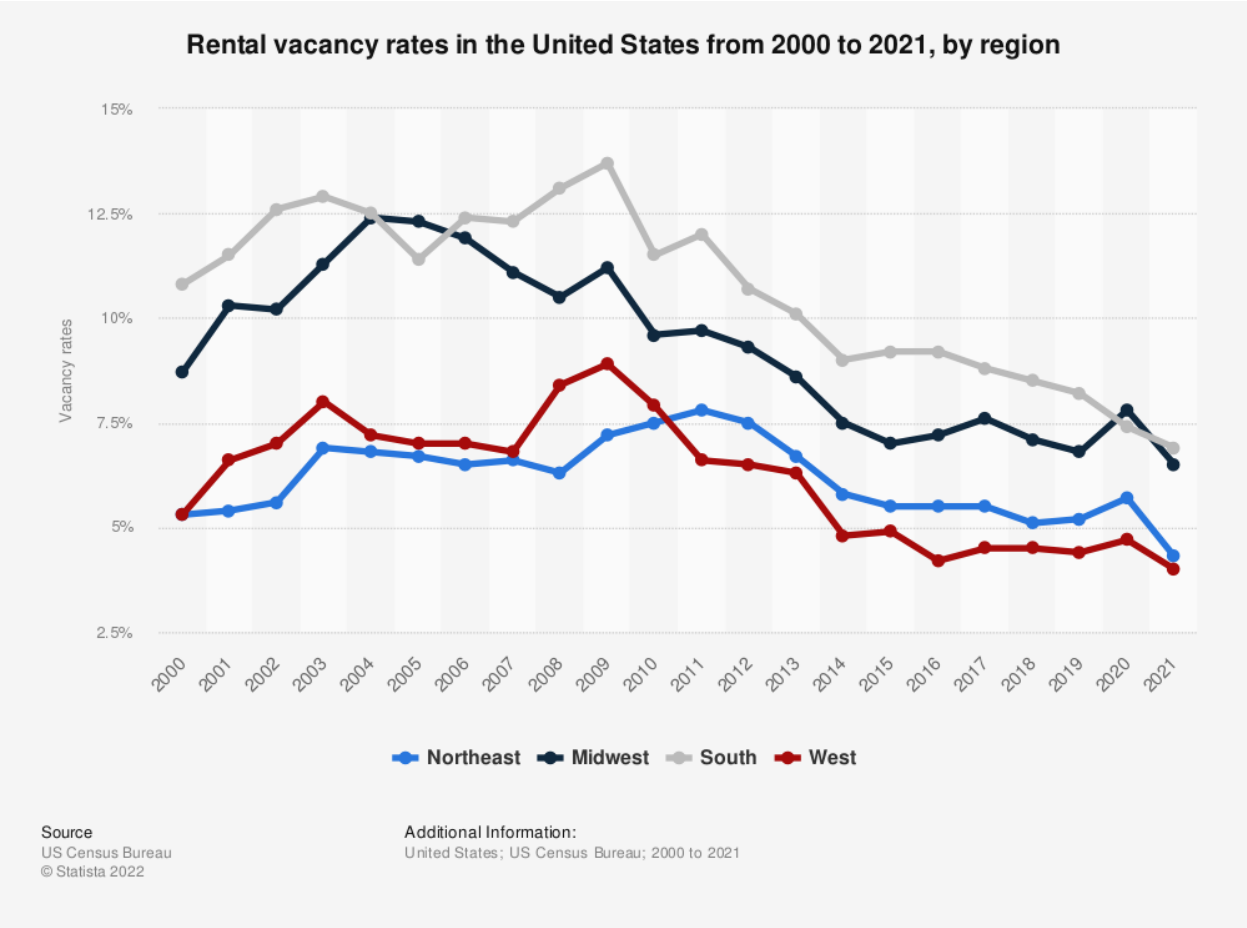



Figure 1: A caption

## 1.2 Description of Data

The data is gathered from Kaggle, a huge repository of community published code and data. This data was pulled from Rentler.com on 7/12/2021, 8/12/2021, and 9/6/2021, and population density data was scraped by zip code from mapszipcode.com on 7/12/2021. The pull from Rentler.com resulted in 4 CSV files which included the main rental listing, the list of amenities, the list of lease terms, and a list of who was responsible to pay each utility. Many of the variables that were sparsely populated were dropped before denormalizing the dataset. The rental listing information was joined with the population and population density information from mapszipcode.com (Source: Kaggle). Many of the data columns in the dataset are embedded in binary format with 0 representing the absence of the predictor attribute while 1, shows that the attribute is present. The data file is massive at around half a Gigabyte. Working with this size of data, we would have to use EDA or take a subset of the dataset for R to run effectively and not crash over the large size of the data. We will explore this idea in later sections. For now, the data is sufficient for analysis.

Our response variable is PRICE which represents the monthly price for the particular summer listing on rental.com.

Table 1: Variables and their descriptions

| Variable | Description |
|---|---|
| V1 | Numbering for the house number we are looking at (form of house identification) |
| pool | Binary data to show if pool exists in this house (1) or not (0) |
| dishwasher | Binary data to show if dishwasher exists in the house (1) or not (0) |
| washer-dryer | Binary data to show if washer-dryer exists in the house(1) or not (0) |
| ac | Binary data to show if air conditioning exists in the house (1) or not (0) |
| parking | Binary data to show if Parking exists in the house (1) or not (0) |
| zip | Zip code of the property |
| price | Monthly rent price for the property |
| city | City where the property is located |
| num_beds | Number of beds in the property |
| num_baths | Number of baths in the property |
| house_type | Type of house we are looking at |
| sqft | Square Feet in the property |
| smoking_ind | Does the rental allow smoking (Yes/No) |
| pets_ind | Does the rental allow pets (Yes/No) |
| acres | Number of acres rental includes |
| description | 4000 character listing description of the rental |
| ZipCity | Primary city for the zip code |
| Population | Population in the zip code |
| PopulationDensity | Population density per square mile for zip code |
| security_deposit | Security deposit required |

## 1.3 Goal

The goal of this study is to analyze the most important factors affecting the housing prices in the US. We will be using the variables in the dataset to do so. In our preliminary market analysis, we found that housing prices are determined by many factors and hence we will try to ascertain factors which have a significant impact on housing prices. Our goal is to build a model that will give us the value added or subtracted from a house with/without the presence of a variable factor. This observation will benefit people who are looking to rent properties in the US and can help them get a better value for the kind of place they may be looking for.

## 1.4 Summary of Findings

### 1.5 Issues and Limitations

The biggest issue we initially faced was with respect to the file size which turned out to be quiet massive even for R Studio. The raw data we started with was half a gigabyte big which turned out to be very massive for any for of extrapolation. Hence we had to shorten the data out. . . . . .

## 2 Exploratory Data Analysis

### 2.1 Data Preprocessing/Cleaning

#### 2.1.1 Read the data

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

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

#### 2.1.2 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

- link
- street_address
- full_address
- acres
- description

```
data_20000_filter <-
  data_20000 %>%
  select(-link, -street_address, -full_address, -acres, -description)
```

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

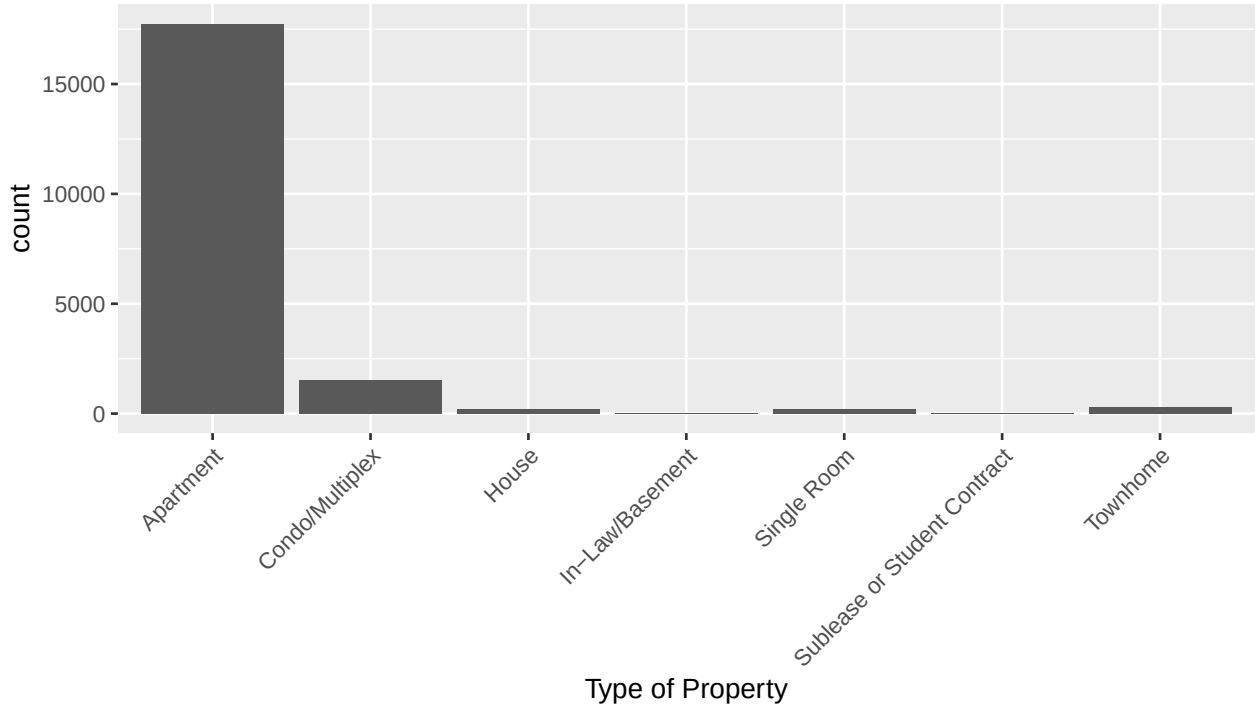
```
data_20000_filter[is.na(data_20000_filter)] <- 0
summary(data_20000_filter)

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)
}
```

## 2.2 Data Transformations and Plots

```
p<-ggplot(data=data_20000_filter, aes(x=house_type)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  xlab('Type of Property') +
  geom_bar()
p
```



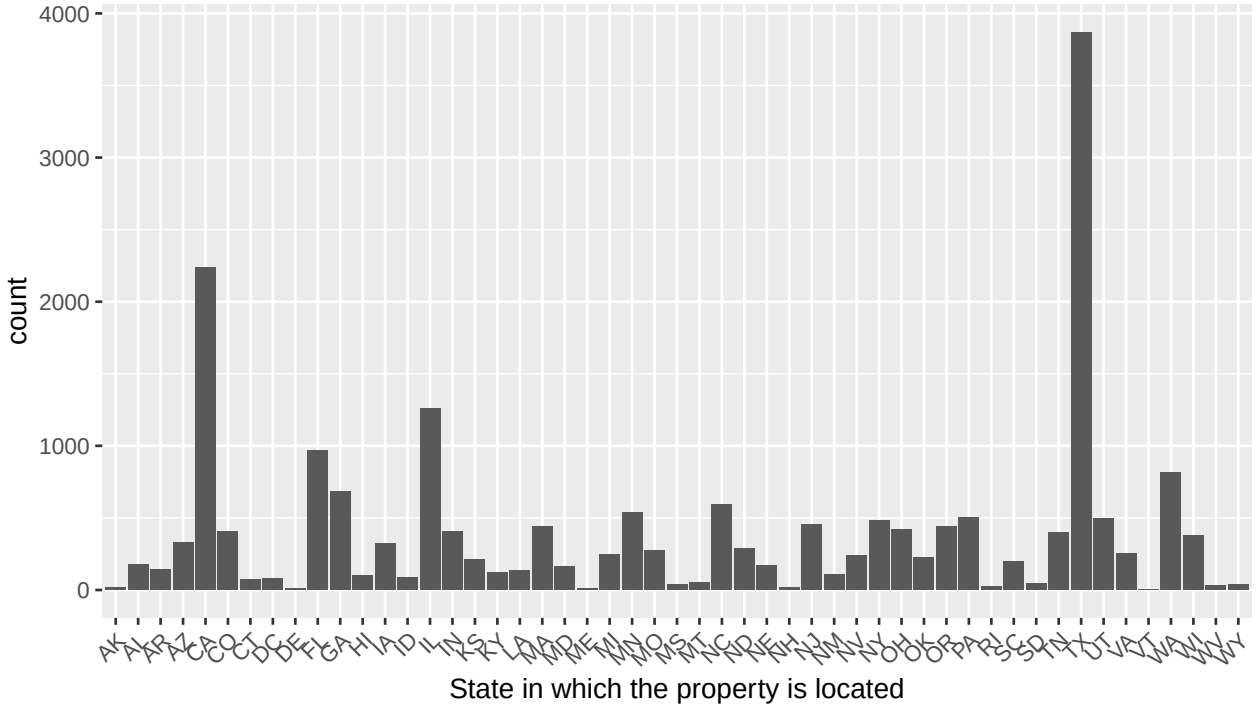

```
y <- count(data_20000_filter, house_type)
print(y)
```

```
                house_type     n
```

1: Apartment 17749 2: Condo/Multiplex 1519 3: House 219 4: In-Law/Basement 26 5: Single Room 211 6: Sublease or Student Contract 1 7: Townhome 275 Looking at the data we see that the properties we will most be evaluating will be Apartment style places with 17749 observations. Other places are lesser in number but still there. The second largest group is the Condo/Multiplex group that we are looking at with 1519 observations. Other than that the smallest group we see is the sublease or student contract group which only has 1 observation.

```
p<-ggplot(data=data_20000_filter, aes(x=state)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  xlab('State in which the property is located') +
  geom_bar()
```

```
p
```




```
number_states <- data_20000_filter %>%
  group_by(state) %>% count()

arrange(number_states, -n)
```

The data represents all states, some more than others. Texas is the most represented state with the least being Vermont at 4 listings. The sample is representative of all states in the US. We are randomly choosing 20,000 data points so this will fluctuate if we change the data seed.

To do the further model analysis, we transform some of the data. Firstly, convert the following columns from characters to binary data

- house_type
- smoking_ind
- pets_ind

```
data_model <- data_20000_filter
data_model$house_type <- as.numeric(as.factor(data_model$house_type))
data_model$smoking_ind <- as.numeric(as.factor(data_model$smoking_ind))
data_model$pets_ind <- as.numeric(as.factor(data_model$pets_ind))
```

```
data_model
```

# 3 Model Training

Firstly, we split the data for choosing and validating and model

- train: 13000
- test: 5000
- validation: the rest

```r
set.seed(1)  # for the purpose of reporducibility
n <- nrow(data_model)
train_test.index <- sample(n, 18000)
train.index <- sample(train_test.index, 13000)
#train.index <- sample(n, 13000)
test.index <- sample(n, 19999) - (sample(n, 19999)-train_test.index) - train.index

# Split the data

n1 <- floor(13000)
n2 <- floor(5000)
set.seed(1)
idx_train <- sample(n, n1)
idx_no_train <- which(! seq(1:n) %in% idx_train)
idx_test <- sample(idx_no_train, n2)
idx_val <- which(! idx_no_train %in% idx_test)

data.w.price.train <- data_model[idx_train,]
data.w.price.test <- data_model[idx_test,]
data.w.price.val <- data_model[idx_val,]
```

# 4 Performance Analysis

# 5 Conclusion