Extracting Features using Regex and stringr package of Messy(unstructured) Craigslist House Listings data set

Ankita Giri

09/03/2021

Data:

The data set used for analysis in this project are posts for rental apartment listings on Craigslist. It is an unstructured data with 601 text files and each file represents each apartment listing posts made on Craigslist. The dataset and rmarkdown can be accessed from the above Code access link. The location of the listing posts are California, USA. The data set is used for extracting several information about rental prices, deposit amounts, kinds of pets allowed, pet deposit, heating and air conditioning systems. The information is further used to analyze any trend or patterns.

This project was done as a part of project submission in a class.

`read_posts` and `read_all_posts`:

Two functions read_posts and read_all_posts are created to control loading and usage of the data set. The function read_posts loads files. This function has only one parameter called "file" for controlling which file is loaded. As an addition to the suggested function format, Using str_c() the contents in each file is combined as a single character vector for convenience. The function read_all_posts loads directories. The function has a parameter "directory" that controls which directory is to be loaded. The previously created function read_posts is used with in the function read_all_posts to control the files that are loaded within the directory that read_all_posts loads. The function read_all_posts returning value is a data frame where the columns is a collection of files and each row is a character vector (each file represents a single character vector). In this project, the "housing" directory is the data set used and the files are each listing posts in the mention in the "housing" data set.

Rental Price

The title for each post is on the first line. From the combined string for each post, the first `\n` will represent the end of the first line which is the title for the post. Then, the separated title and the remainder of the information is stored as two separate columns as a data frame called "posts". From the title, the amount

Table 1: House Listings

	Listings
housing/7356424805.txt	$2,295 / 2br - 680ft2 - $2295-2beds and 1 bath in center sunset (sunset / parkside) viewport: w
housing/7356680852.txt	$3,080 / 2br - SPACIOUS 2 BEDROOM IDEAL SF LOCATION WITH LAUNDRY!! (pacific h

Table 2: House Listings and Rental Price Extracted(first few)

	Listing
housing/7356424805.txt	$2,295 / 2br - 680ft2 - $2295-2beds and 1 bath in center sunset (sunset / parkside) viewport: w
housing/7356680852.txt	$3,080 / 2br - SPACIOUS 2 BEDROOM IDEAL SF LOCATION WITH LAUNDRY!! (pacific h

Table 3: Number of Listings with and without Rental Price

No Listing Price	Number of Listings
FALSE	598
TRUE	3

value of the rental price for the apartment listing is then extracted and stored in a new column as a numeric value.

Using `is.na()`, a check for the NA values is done to make sure that the majority of the listing prices are extracted. There 3 listing which does not have rental price included in the listing.

Limitations:

There are 3 posts that do not have listing prices on the title so they are returned as NA values.

Also, some posts contain additional listing for more apartments that is not in the title. Therefore such listings have not been accounted for. So, the price column do not accurately represent all the apartment listings in the housing directory.

Deposit Amount:

The deposit amount is listed in different patterns in the listings so several patterns for extracting the deposit amount is explored.

Some patterns explored are:

- Deposit: $amount
- Deposit $amount
- Deposit($amount)
- Deposit (one word) $amount.
- Deposit (two words) $amount.
- $amount Deposit
- $amount.0000 Deposit
- $amount (one word) Deposit
- $amount.0000 (one word)Deposit
- Deposit: $amount

Using regex patterns and string manipulation functions from stringr package, the deposit amounts are extraced and all the extracted values are stored in the column deposit_amount in the housing data frame.

Some Limitations of Extracting Deposit Amounts:

Table 4: House Listings and Deposit Price Extracted(first few)

	Listing
housing/7356424805.txt	$2,295 / 2br - 680ft2 - $2295-2beds and 1 bath in center sunset (sunset / parkside) viewport: w
housing/7356680852.txt	$3,080 / 2br - SPACIOUS 2 BEDROOM IDEAL SF LOCATION WITH LAUNDRY!! (pacific h
housing/7356706648.txt	$1,695 Spacious Studio in Great Location! By Trader Joe's (nob hill) viewport: width=device-w
housing/7356706929.txt	$2,495 / 2br - 2BR in Nob Hill; Convenient Downtown Location! (lower nob hill) viewport: wid
housing/7356710385.txt	$1,595 Studio Available; HWF; Market and Gough (hayes valley) viewport: width=device-width

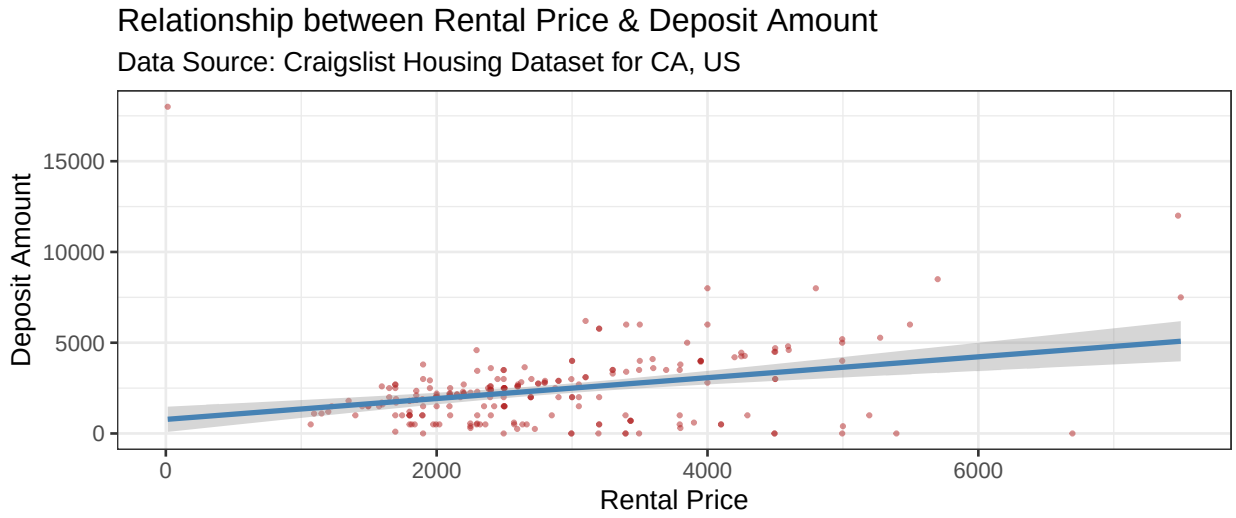
Table 5: Number of Listings with and without Deposit Price

No Deposit Price	Number of Listings
FALSE	221
TRUE	380

- Posts with multiple listings had multiple deposit amounts, hence only the first pattern match is extracted.
- There are several unextracted pattern formats like "Deposit: 1 month's rent", "Deposit: 1.5 * rent", "Deposit: One month rent" for the purpose of keeping the values numeric. This was done so that the relationship between the rental price and the deposit amount could be explored.
- It is also possible that unknown patterns may not have been explored.
- Some posts may have listing for pet deposits only and hence the extracted deposit amounts may have included pet deposit listings too.

There are 380 listings whose deposit feature either did not exist or has not been extracted because of the limitations mentioned earlier.

Relationship between Rental Price and Deposit Amount:



The linear regression line is fitted in the scatter plot between the rental price and the deposit amount shows that there is a slight positive correlation between the two. This means that for most apartment listings, the higher the rental price, the higher will be the deposit amount.

Table 6: House Listings and Pets Allowed Information(first few)

	Listing
housing/7356424805.txt	$2,295 / 2br - 680ft2 - $2295-2beds and 1 bath in center sunset (sunset / parkside) viewport: w
housing/7356680852.txt	$3,080 / 2br - SPACIOUS 2 BEDROOM IDEAL SF LOCATION WITH LAUNDRY!! (pacific h
housing/7356706648.txt	$1,695 Spacious Studio in Great Location! By Trader Joe's (nob hill) viewport: width=device-w
housing/7356706929.txt	$2,495 / 2br - 2BR in Nob Hill; Convenient Downtown Location! (lower nob hill) viewport: wid
housing/7356710385.txt	$1,595 Studio Available; HWF; Market and Gough (hayes valley) viewport: width=device-width

Table 7: Number of Listings without Pet Information

No pets info extracted	Number of Listings
FALSE	446
TRUE	155

Pets:

Categorical Feature that measures whether the apartment allows pets: cats, dogs or both, or none and also other kind of pets if any.

For the Categorical feature measuring whether the apartments allow pets, information if cats, dogs or both are allowed is relevant. Also information regarding whether if no pets are allowed is relevant.

If the type of pet is not mentioned, the value is "yes". The rest is NA values due to lack of information.

Patterns tested for pets allowed:

- Pet- friendly
- Pets are OK
- Pets are allowed

Patterns tested for pets not allowed:

- No pets allowed
- Pets not allowed
- Pets are not allowed

Patterns tested for cats:

- Cats are OK

Patterns tested for Dogs:

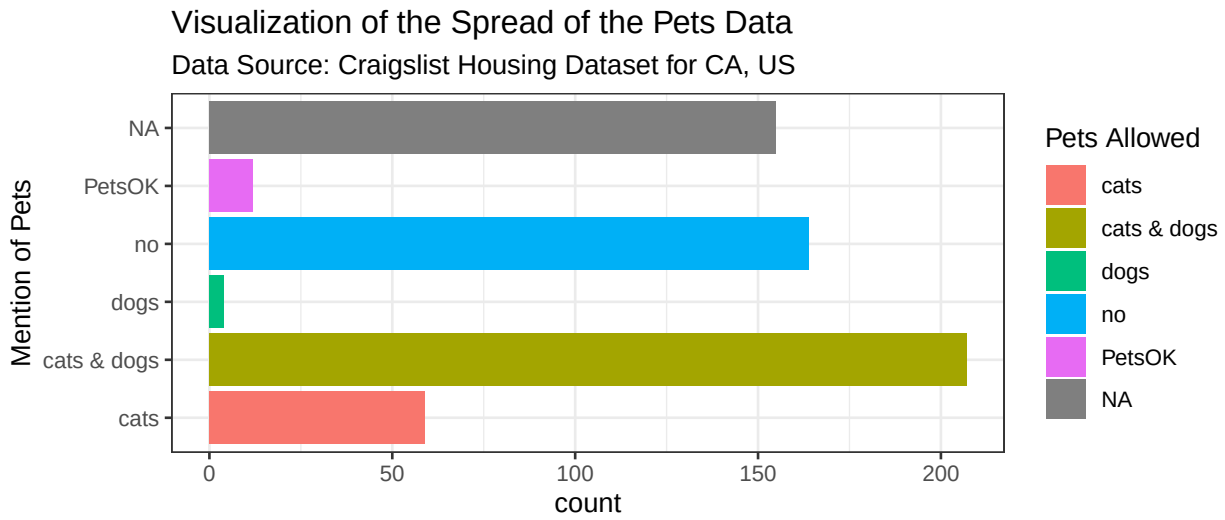
*Dogs are OK

A pattern check was done to check whether other pets are allowed and no conclusive information was found.

There are 155 listings with no pet information. This could have occurred because of unexplored patterns for some category.

Table 8: Number of Listings without Pet Deposit Information

No pets deposit information	Number of Listings
FALSE	6
TRUE	595



Limitations:

For the Pet Deposit, the data is extremely sparse therefore has been ommitted from visualization. Upon exploring several reasons were found like:

- Pet Rent is taken instead of Deposits
- All existing Patterns for Pet Deposit may not have been explored
- Another existing key word was Pet Fee for each cat and dog that is charged monthly

Heating and Air Conditioning

The heating system are categorized into different factors like fireplace, heater.

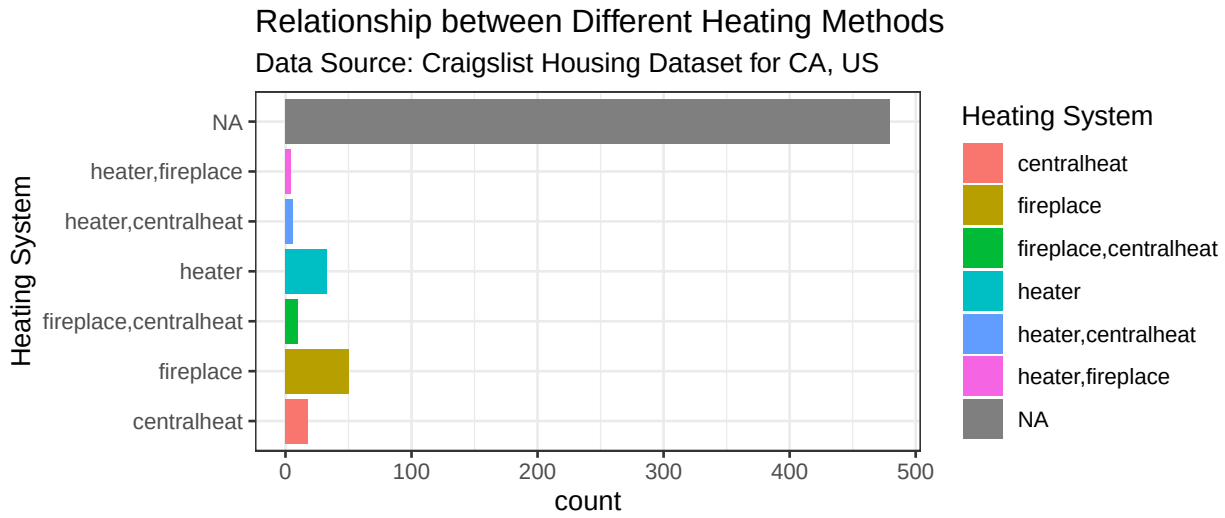
Below is a graphical representation of the spread of the data.

Table 9: House Listings and Available Heating System

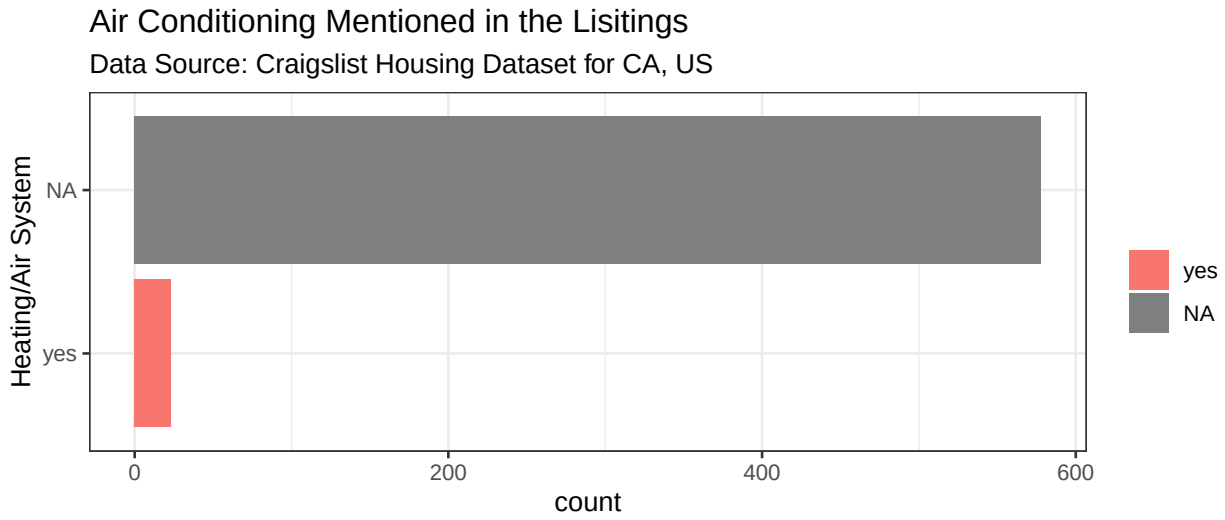
	Listing
housing/7370168508.txt	$2,395 / 1br - 790ft2 - Available now 1x1.. Grand 1x1 In Santa Clara!!Call now.. (santa clara) v
housing/7370168640.txt	$2,149 / 2br - Coming Available In September! This 2x1.5 Won't Last! Plan NOW! (santa rosa)

Table 10: Number of Listings without Heating System Information

No Heating System information	Number of Listings
FALSE	121
TRUE	480



There are 390 listings without heating system information.



Limitation:

There are several missing values which could indicate either relevant patterns are not explored or that the information regarding heating is not included in the listing post.