Sta 141a

Rohan Malhotra
Statistics Department
University of California, Davis
December 2018

As discussed in the previous assignment, Craigslist is a website, in which one can easily post advertisements on what product they are selling. The last assignment consisted with the analysis of the spatial characteristics of the Craigslist data set, which consisted of apartment rental listings in California. This assignment deals, with scraping the data. In the previous, assignments the data of all the features from each craigslist was given to us. The goal of this assignment was to clean and extract those features from each craigslist post, in order to create a data set similar to the one in previous assignments.

1 Creating the data set

First, since there is obviously no function, that can read a text file, and extract the desired features. Therefore, I created two functions to create accomplish this task. These two functions were read_post and read_all_posts. The first function, read_post takes a text file as the argument and reads in the lines of the text files. The second function, read_all_posts, takes in a directory (craigslist posts of a certain city) as the argument, and will use the read_post function to read each text file within the directory and then extracts certain features from the file. When doing this, the read_post function was modified so it reads all the lines of a text file, but also puts the whole text file into one line, which makes text processing more convenient. Each row consisted of an apartment listing and the columns (features) include: title of the post, text of the post, number of bedrooms and bathrooms, sqft, upload date of the post, and the price of each apartment. These are the columns of the data frame, and were specifically chosen because they were easy to extract as each post had the same pattern for each of these features.

2 Extracting more complex features

The basic features mentioned above were easy to extract because each post had the same pattern for each feature. These upcoming features that were extracted do not have the same pattern for each post. Therefore, as an analyst one must accept the fact that it is impossible to get 100% of the data. Therefore, the goal was to find the most common patterns that occurred in each post description for each feature and extract those. Additionally, instead of one directory, or one city, all the posts from each city were used, which came out to be 45,845 craigslist apartment postings.

2.1 Title price versus user price

Craigslist, asks the user to input the price in a separate field when the post is created, this is known as the user price. Also, users tend to put the price in the title of their post, to make it easier for potential buyers to see the price without opening the whole post. Before, I compared these two price, I expected them to be the same for the most part. In order to, test the hypothesis I extracted both title and user price. From this, there were 176 posts that did not mention the title price and

180 posts that did not mention the user price, below is a table which illustrates the frequencies of the absolute value of differences between the user and title price.

|user price – title price|	0	50	70	100	200
Frequencies	45,653	3	2	3	4

From the table, above it is seen, that primarily the title price matches the user price, but sometimes users tend to either charge a little more or accidentally press the wrong number when typing in the user price.

2.2 Apartment price versus security deposit price

Next, I wanted to see the relationship among the user price and the security deposit price. My prediction is that they have a positive linear relationship, meaning as the apartment price rises so does the security deposit price. Before, comparing them extracting the deposit price was tricky as a post consisted of many other deposits such as pets or moving. Therefore, after looking at the frequencies of the different deposits, I excluded the most frequent non-security deposit prices, but keep in mind not all non-security deposit prices were excluded. The figure bellows shows a plot of apartment prices vs. security deposit prices. Before graphing, I made some reasonable assumptions to only look at posts with an apartment price of above $200 and security deposit prices above $0 to get the most accurate representation.

From the figure, above, there is a linear relationship present, which confirms my hypothesis. Furthermore, the correlation coefficient is 0.59 meaning that it there is some what a relationship among apartment prices and security deposit prices.

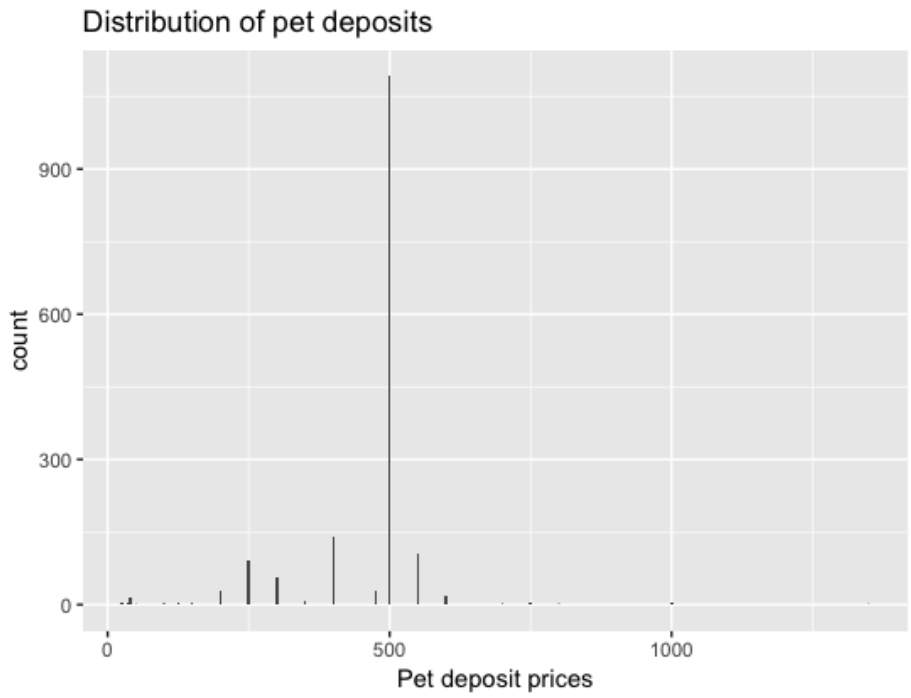
2.3 Pet policy and deposits

Next, I was interested in extracting pet policies and deposits. First, the most common pets are dogs and cats therefore I looked at the distribution of the pet policy among these two pets. Obviously, this was challenging as there were so many different patterns for dogs only, cats only, both or none so I extracted these features based off the most common patterns I notices when looking through the posts. Below shows the distribution of these pet policy.

Pet Policy	Dogs only	Cats only	Both	None
Frequency	725	1,519	11,247	5,484

The table, displays that most apartments tend to allow both cats and dogs. Also I conclude that most people prefer cats over dogs as there are almost the twice amount of cats as there are dogs. Furthermore, besides the most common pets some posts also allowed other pets such as: birds, hamsters, gerbils, rabbits, guinea pigs, chinchillas and aquarium/terrarium animals including fish, hermit crabs, turtles, frogs, and small lizards. People tend to allow small caged animals and mention that they do not tend to accept exotic pets.

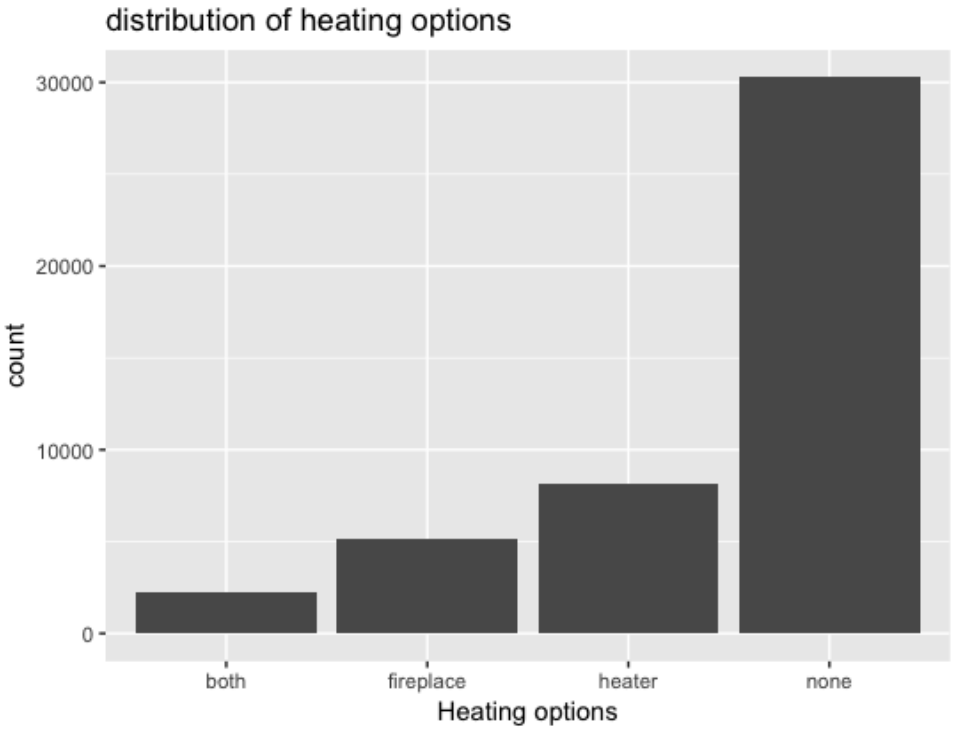
Next, I also looked at the distribution of just the pet deposits. Again, when extracting, there were many different patterns for pet deposit as some mentioned different prices for each pet, or had different rates so the data is not 100 % accurate. The figure below shows the distribution of pet deposits.



The figure above clearly shows that the distribution is skewed, and that for the most part pet deposit prices tend to cost around $500 as it is the most frequent occurring price as shown by the figure.

2.4 Heating and cooling options

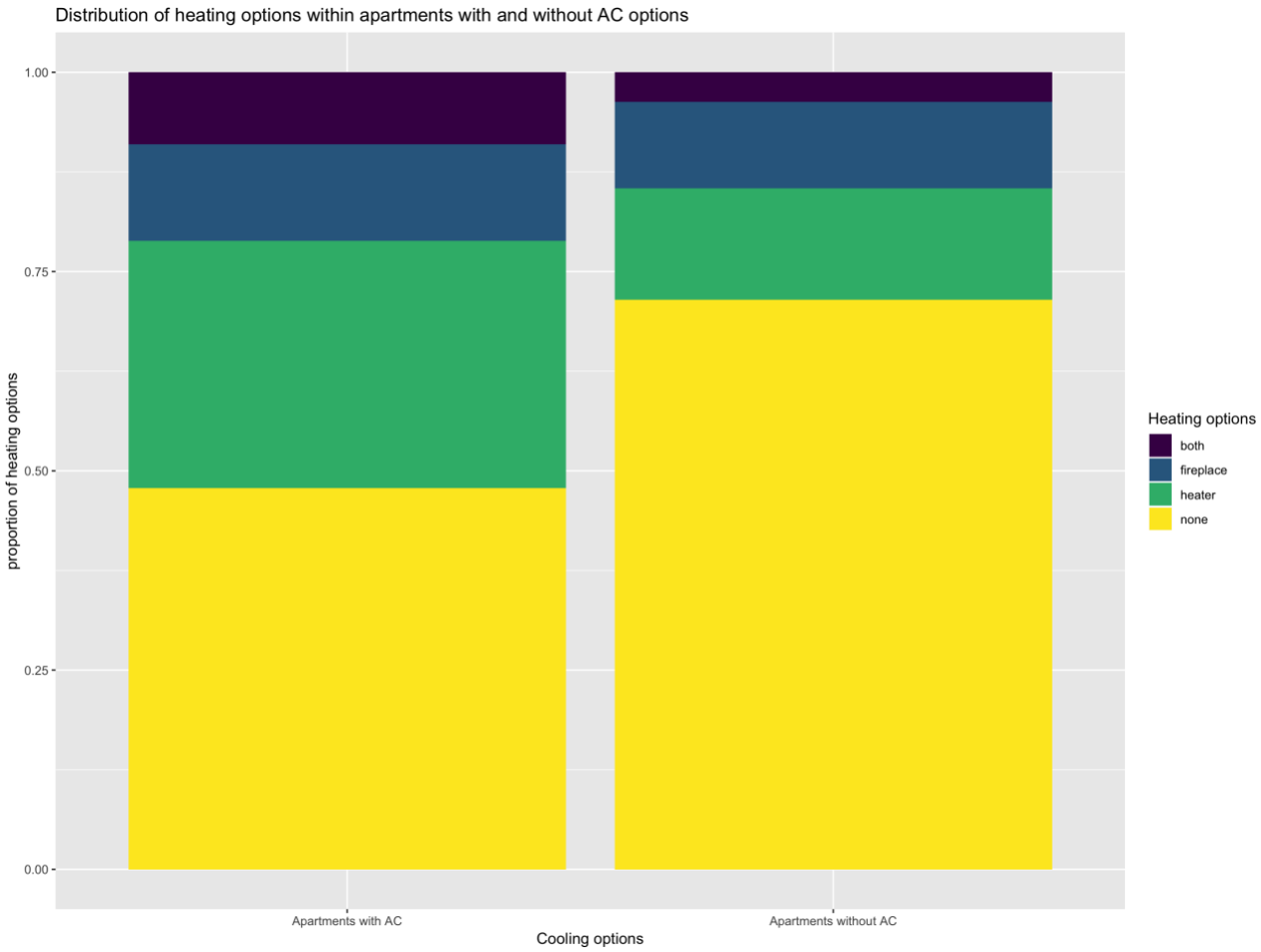
Next, looking at the heating options, apartments tend to have either a fireplace, heater or both. Below shows the distribution of these heating options.



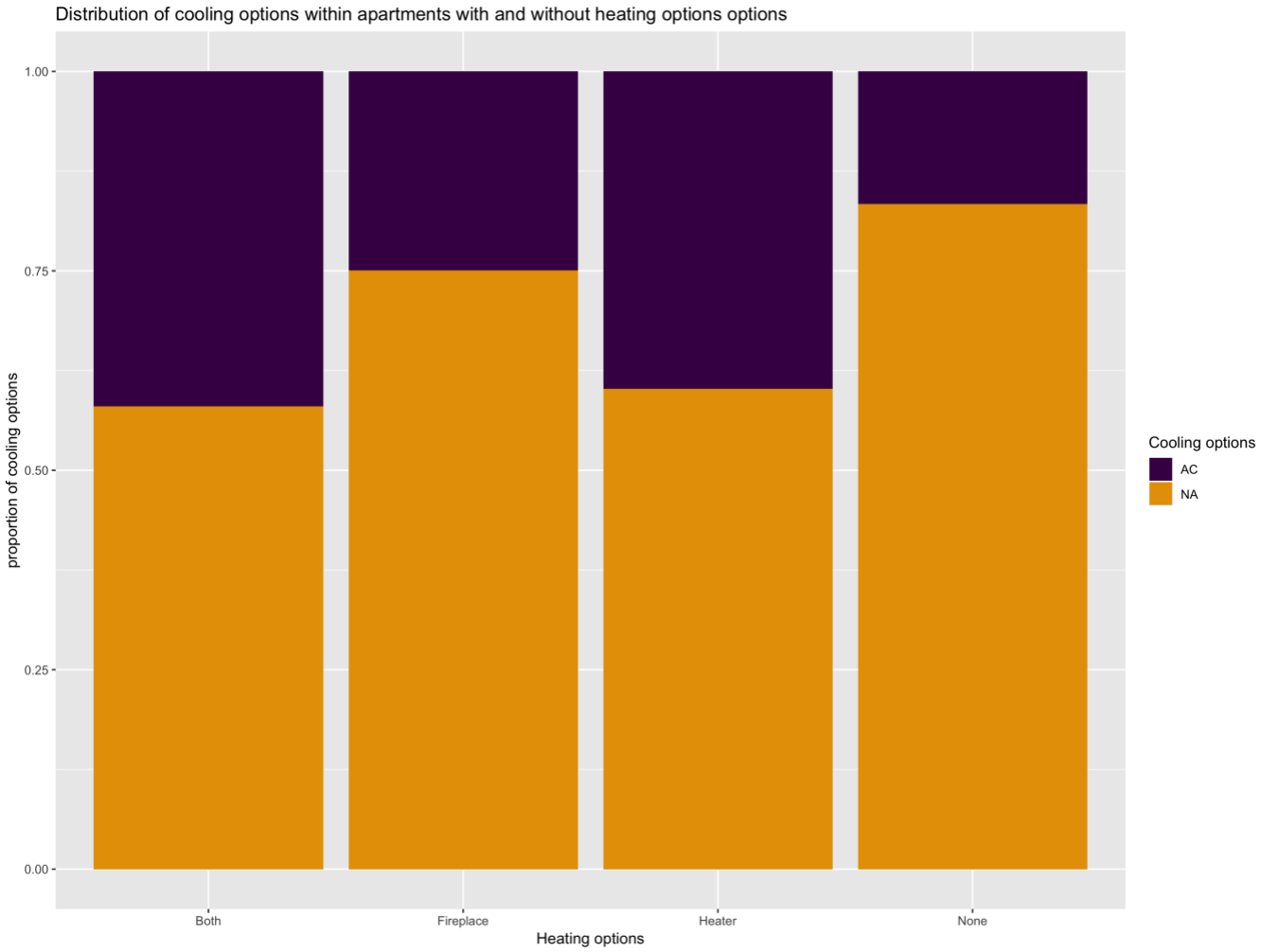
It is seen that most apartments tend not to offer heating options, but the most common heating option is the heater. Additionally, apartments tend to offer one or the other, but do not offer both very often.

For cooling options, there is only an air conditioner (AC). Overall, only 10,521 apartments offered A/C where as 15,560 apartments offered a heating option. Therefore, it seems that heating tends to be more commonly offered than cooling.

Also, I am to curios to see if there is some sort of relation among the cooling and heating options, so I will first look to see do apartments with air conditioning tend to have heating. This can be depicted in the visual on the following page.



The graphic displays the proportion of heating options among, apartments with an AC and without an AC. It is seen apartments that have a cooling option tend to having a heating option as 50% of apartments with an AC have some sort of heating option. On the other hand, apartments with no cooling option tend not to have a heating option as well. Next, I looked at if apartments with a heating option tend to have a cooling option. This can be seen from the figure on the next page.



The figure above shows the proportion of cooling options among apartments with a heating option. From this, I conclude typically apartments that have a heater or both a heater and a fireplace will tend to have an AC as well.

2.5 Privacy feature

Craigslist, gives people an option to hide their email and phone number in the description. They offer this feature to prevent web scrapers to get that privacy information. The goal is to see if people make use of that feature. Both emails and phone number have a pre specified pattern so that is what I used to extract those features. First, looking at emails, only 14 emails were found. Next, looking through phone numbers there were 34 phone numbers found. Finally, 34,800 posts had “show contact info” option which means they had that feature on. From this though there will be some posts that did not have a phone number or email or “show contact info”, I believe I have enough data to conclude that people tend to have the feature on because I found over 60% of people did have that feature on.

3 Conclusion

In conclusion, the way I extracted the basic features from the read_all_posts function I feel like was the best way as I could use that function for further string processing which is seen in section 2 of the paper. I believe that within the data set though I could not account every single pattern I could get valuable insight by extracting the most frequent patterns for each problem. This assignment taught me the more harder job of a data analyst as in the real world data will not be given, and extraction methods used in the assignment are crucial to get the data which could further be analyzed.