COLUMBIA UNIVERSITY
School of Professional Studies

Cleaning and Processing Dataset of Airbnb Property from Texas

by

Wei Yu

Project for the Course of Applied Analytics Frameworks & Methods
Instructor: Vishal Lala
Facilitator: Xu Zoe Zhu

March 1st, 2018

Contents

Chapter 1

Introduction

Airbnb, founded in 2008, is a private US company that offers a online platform for landlord to rent their houses and/or rooms to visitors who are traveling to those cities. For travelers, not only Airbnb provides them with millions of affordable housing choices and accommodations, but also provides unique travel experience with local culture. For local communities, they are able to increase their income by renting out extra unused spaces to visitors.[1] This industry has grown gradually all over the world and there are now over 4 millions listing in 65,000 cities and 191 countries.[2]

Our group has found some Airbnb listing data in Texas, US. The data is downloaded from Kaggle, which is a platform that contains lots of source data for data mining. The website address which can be used to download the source data is stated in reference [1] The name of the original data is "Airbnb Property Data from Texas", and there are a few reasons we choose this data. Vacation rental is increasingly popular nowadays and this dataset comprise a representative view to the vacation accommodation market. The Dataset offers comprehensive and practical data for researching the price fluctuations and the popular locations. Additionally, the dataset provides a reliable source of data from the Texas households. With the large amount of data provided by the dataset, it enables further analysis into the Texas rental market, and the variations to the market share of Airbnb in Texas.

1.1 Data Content

The "Airbnb Property Data from Texas" dataset[3] includes more than 18,000 property listings from Texas, US. There are 10 headings in the dataset: Average Rate Per Night, Bedrooms Count, City, Date of Listing, Description, Latitude, Longitude, Title, Property description, and URL.

The data fields “average rate per night”, “bedrooms count”, “city”, “date of listing” are useful and applicable to our research. These four fields of data composes the research variables, helping us to analyze and potentially provides a solution to the research questions. Below are the formed research questions:

1 How does the distribution of properties that are listed on Airbnb change in each city of Texas?

2 What is the spread of pricing to the Airbnb listed properties in varies cities in Texas?

3 What is the trend of the listings on Airbnb each year?

Chapter 2

Data Problems Diagnosis

We notify some data problems when we roughly read the Excel and here are the key issues of this dataset:

1 Unreadable Code

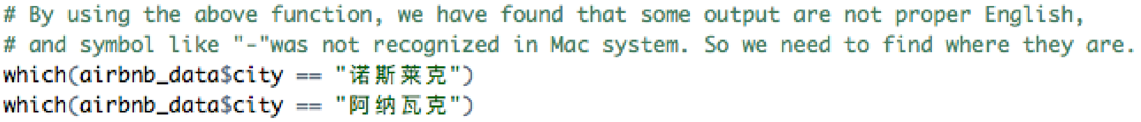
2 Missing Data

3 Unformatted Data

4 Minor Issues that not worth Fixing

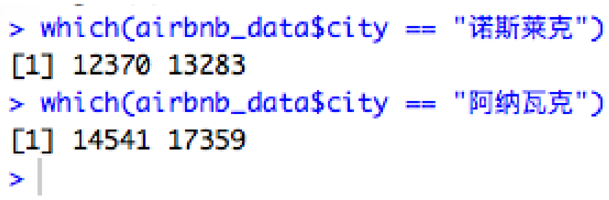
2.1 Unreadable Code

It appears to be some unreadable code in the column of ”city”. The words are in Chinese, whereas they should be written in proper English. To identify and locate when these errors are, use function of ”unique()”. This function helps to identify the weird Chinese output. By using the below R code, the locations of all the weird outputs in the ”city” column are identified.



```
# By using the above function, we have found that some output are not proper English,
# and symbol like "-"was not recognized in Mac system. So we need to find where they are.
which(airbnb_data$city == "诺斯莱克")
which(airbnb_data$city == "阿纳瓦克")
```

As a result, the output shows to be as follows:



```
> which(airbnb_data$city == "诺斯莱克")
[1] 12370 13283
> which(airbnb_data$city == "阿纳瓦克")
[1] 14541 17359
> 
```

N.B: The result of unique() function is too long so it is not attached here.
The result shows that there are two entry of the first Chinese output and another two entry of the second Chinese output.

2.2 Unformatted Data

It is obvious to find out that in the column of "bedrooms_count", there are lots of "numbers" shown as "studios". This can cause significant inconvenience in later analysis of the dataset. As a studio only has a single undivided space as a bedroom and living room, we classify the studios to be equivalent to 0.7 bedrooms apartments. This data is then well transformed and is more suitable from analyzing the overall prices of the Airbnb listed properties.

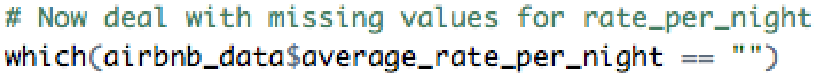
On the other hand, similar problems occur in the column of "average_rate_per_night". All the data in this column is shown as $ plus a number, which is not a proper numeric format in R.

Also, in the "city" column, it appears to be an error when using symbol "-", especially on Mac. This error appears in the entry of "BryanCollege Station". For the convenience of future analysis, we will need to get rid of the symbol "_".

2.3 Missing Data

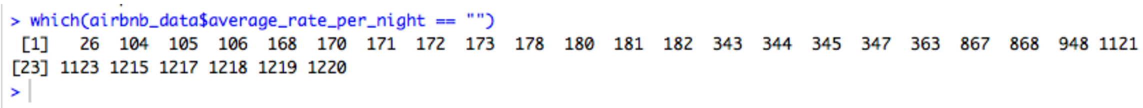
There are several empty entries in this dataset. In this way, we try to use code of "!complete.cases()" to identifying all the missing values as the first step. However, it shows to have too many rows with issues. Instead, we then focus on each specific columns with which data are meaningful to the research.

Firstly, check with "average_rate_per_night" column. This column contains data about the average price per night of each Airbnb property and therefore is the most important column in this dataset. Now use "which" function again to check all the empty entries:



```
# Now deal with missing values for rate_per_night
which(airbnb_data$average_rate_per_night == "")
```

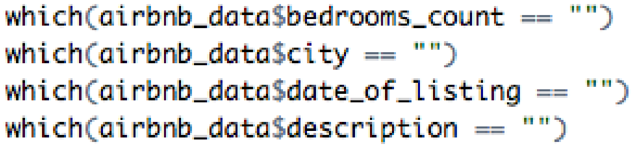
Here is the output of the above function:



```
> which(airbnb_data$average_rate_per_night == "")
 [1]   26  104  105  106  168  170  171  172  173  178  180  181  182  343  344  345  347  363  867  868  948 1121
[23] 1123 1215 1217 1218 1219 1220
> 
```

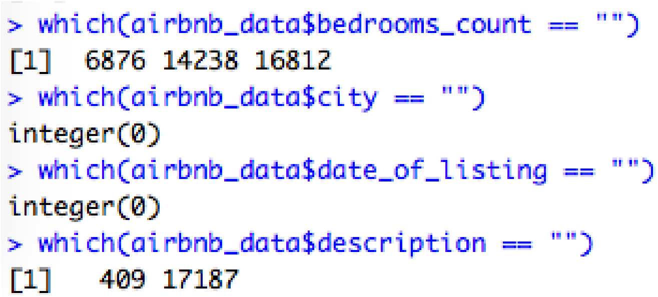
It seems there are quite a lot of empty entries in this column and we need to fix it in the future.

Now turn to other columns such as "bedrooms_count", "city", "date_of_listing" etc. Again use "which" function to check the entries of them:



```
which(airbnb_data$bedrooms_count == "")
which(airbnb_data$city == "")
which(airbnb_data$date_of_listing == "")
which(airbnb_data$description == "")
```

Here is the output of the above function:



```
> which(airbnb_data$bedrooms_count == "")
[1]  6876 14238 16812
> which(airbnb_data$city == "")
integer(0)
> which(airbnb_data$date_of_listing == "")
integer(0)
> which(airbnb_data$description == "")
[1]   409 17187
```

From the result we can observe that there are several properties that do no have number of rooms or descriptions in their dataset. However, all the properties do have data about where they are (i.e. the city) and when the property was listed on the website (i.e. date_of_listing).

We will need to fix them and these will be explained in detail in Chapter 3.

2.4 Minor Issue that not worth Fixing

The fields which are minor and does not have a direct impact to our analysis are ignored. These fields are ”description”, ”latitude”,”longitude”, ”title” and ”url”. There are 34 rows of data for which the data for these fields are missing. We were able to recover the ”latitude” and ”longitude” data by looking up the city online. Even though we have recovered the missing data for consistency purpose of the dataset, these field is redundant and can be derived from other critical fields, for example, the ”latitude” and the ”longitude” fields.

There are some missing, unreadable and unformatted data in ”description” and ”title”, but those data does not have a correlation or help to answering the previously defined research questions and the aforementioned four research variables.

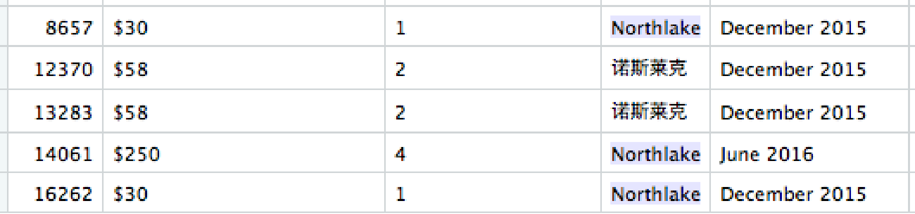
Chapter 3

Dataset Processing

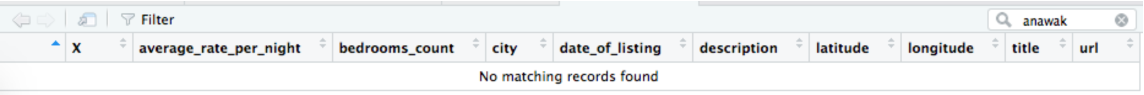
As all the main issues have been identified, we now start cleaning and processing the dataset.

3.1 Unreadable Code

In the chapter 2.1, we have mentioned that there are several unreadable data in the table which need to be fixed. Firstly, use a translation tool to find out the proper English. The two unreadable words should be "Northlake" and "Anawak" respectively. Here are the screen shots before transforming:



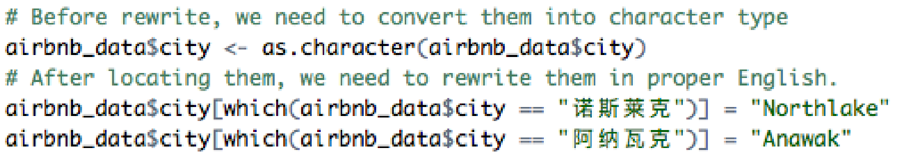
8657	$30	1	Northlake	December 2015
12370	$58	2	诺斯莱克	December 2015
13283	$58	2	诺斯莱克	December 2015
14061	$250	4	Northlake	June 2016
16262	$30	1	Northlake	December 2015



X	average_rate_per_night	bedrooms_count	city	date_of_listing	description	latitude	longitude	title	url

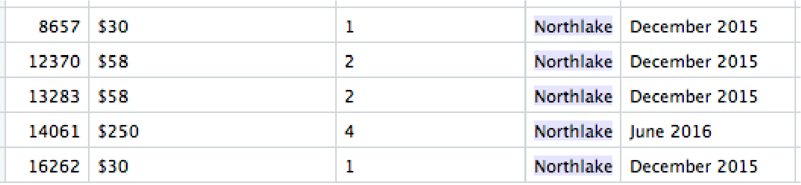
No matching records found

Before transforming, to reduce the risk of unformatted, it is also necessary to use the function of as.character to identify the original format. Thus, by using the following substitution function, we can transform the original text:

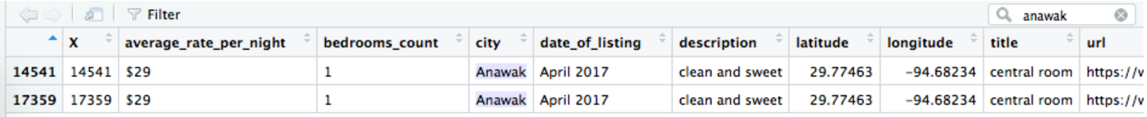


```
# Before rewrite, we need to convert them into character type
airbnb_data$city <- as.character(airbnb_data$city)
# After locating them, we need to rewrite them in proper English.
airbnb_data$city[which(airbnb_data$city == "诺斯莱克")] = "Northlake"
airbnb_data$city[which(airbnb_data$city == "阿纳瓦克")] = "Anawak"
```

After running the above code, check again with these two words:



8657	$30	1	Northlake	December 2015
12370	$58	2	Northlake	December 2015
13283	$58	2	Northlake	December 2015
14061	$250	4	Northlake	June 2016
16262	$30	1	Northlake	December 2015

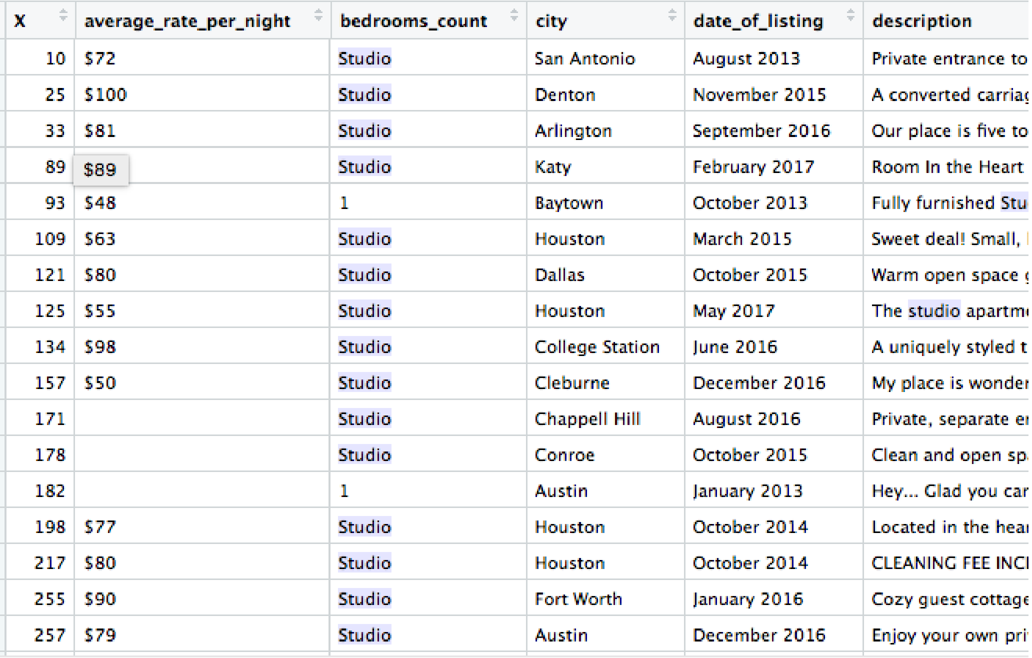


	X	average_rate_per_night	bedrooms_count	city	date_of_listing	description	latitude	longitude	title	url
14541	14541	$29	1	Anawak	April 2017	clean and sweet	29.77463	-94.68234	central room	https://v
17359	17359	$29	1	Anawak	April 2017	clean and sweet	29.77463	-94.68234	central room	https://v

3.2 Unformatted Data

3.2.1 Transformation of ”Studios”

In the interest of convenience, we want to transform all the studios to be counted as 0.7 bedrooms in the dataset. Here is the picture of before transforming:

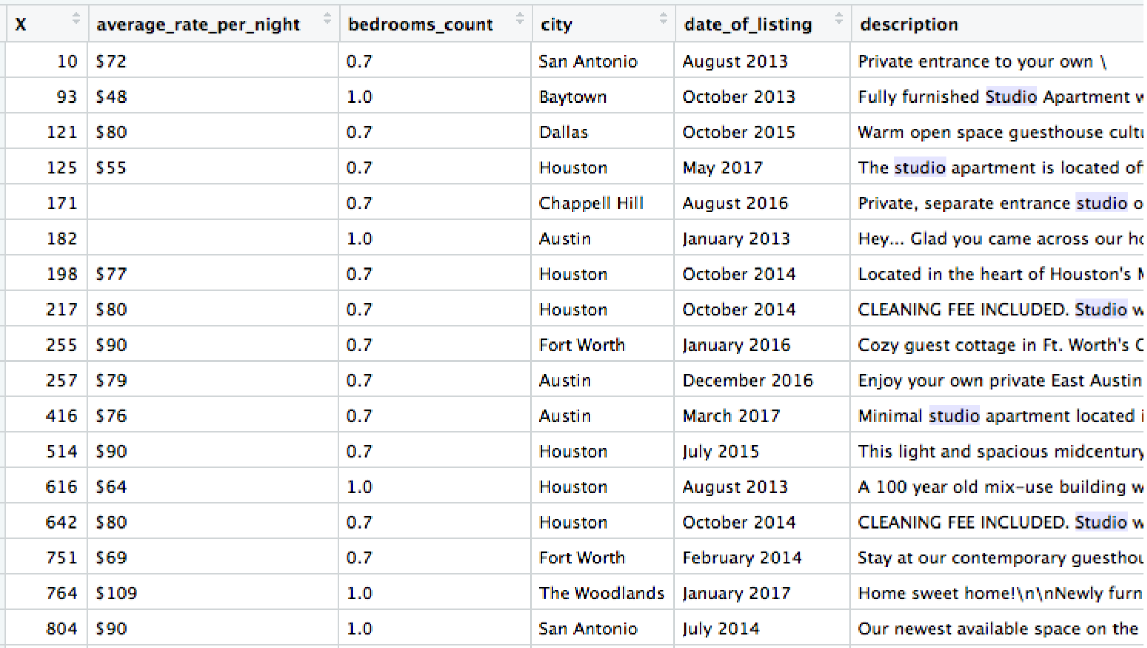


X	average_rate_per_night	bedrooms_count	city	date_of_listing	description
10	$72	Studio	San Antonio	August 2013	Private entrance to
25	$100	Studio	Denton	November 2015	A converted carriag
33	$81	Studio	Arlington	September 2016	Our place is five to
89	$89	Studio	Katy	February 2017	Room In the Heart
93	$48	1	Baytown	October 2013	Fully furnished Stu
109	$63	Studio	Houston	March 2015	Sweet deal! Small,
121	$80	Studio	Dallas	October 2015	Warm open space g
125	$55	Studio	Houston	May 2017	The studio apartme
134	$98	Studio	College Station	June 2016	A uniquely styled t
157	$50	Studio	Cleburne	December 2016	My place is wonder
171		Studio	Chappell Hill	August 2016	Private, separate en
178		Studio	Conroe	October 2015	Clean and open sp
182		1	Austin	January 2013	Hey... Glad you car
198	$77	Studio	Houston	October 2014	Located in the hear
217	$80	Studio	Houston	October 2014	CLEANING FEE INCI
255	$90	Studio	Fort Worth	January 2016	Cozy guest cottage
257	$79	Studio	Austin	December 2016	Enjoy your own pri

Before transforming, it is also worth noticing that "studio" is a word rather than a number. Therefore, to avoid any error, it is recommended to transform all the counts to characters and then back to numeric format. Here are the related code:

```
# In order to find out the studios, first need to change all the counts to "character" type,
# because "Studio" is in "character"
airbnb_data$bedrooms_count <- as.character(airbnb_data$bedrooms_count)
which(airbnb_data$bedrooms_count == "Studio")
# Studios are counted as 0.7 bedroom
airbnb_data$bedrooms_count[which(airbnb_data$bedrooms_count == "Studio")] = "0.7"
# Now change the data type back to numeric
airbnb_data$bedrooms_count <- as.numeric(airbnb_data$bedrooms_count)
```

The result shows to be as follows:

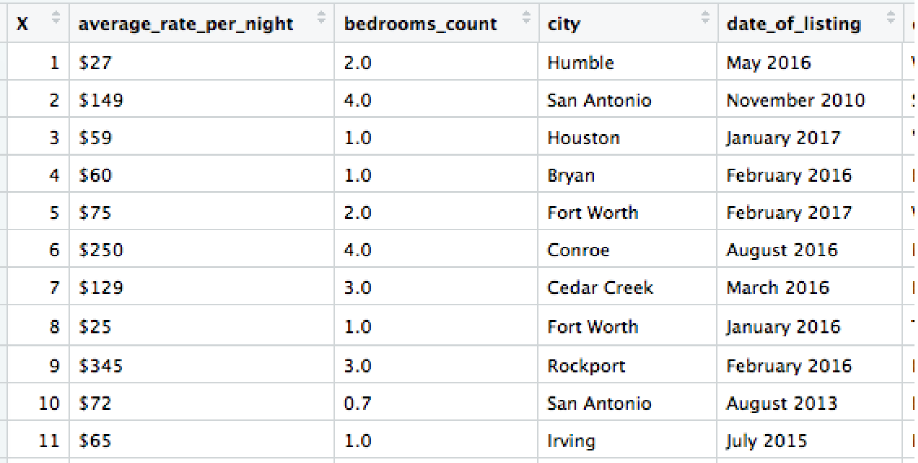


X	average_rate_per_night	bedrooms_count	city	date_of_listing	description
10	$72	0.7	San Antonio	August 2013	Private entrance to your own \
93	$48	1.0	Baytown	October 2013	Fully furnished Studio Apartment w
121	$80	0.7	Dallas	October 2015	Warm open space guesthouse cult
125	$55	0.7	Houston	May 2017	The studio apartment is located of
171		0.7	Chappell Hill	August 2016	Private, separate entrance studio o
182		1.0	Austin	January 2013	Hey... Glad you came across our ho
198	$77	0.7	Houston	October 2014	Located in the heart of Houston's M
217	$80	0.7	Houston	October 2014	CLEANING FEE INCLUDED. Studio w
255	$90	0.7	Fort Worth	January 2016	Cozy guest cottage in Ft. Worth's C
257	$79	0.7	Austin	December 2016	Enjoy your own private East Austin
416	$76	0.7	Austin	March 2017	Minimal studio apartment located i
514	$90	0.7	Houston	July 2015	This light and spacious midcentury
616	$64	1.0	Houston	August 2013	A 100 year old mix-use building w
642	$80	0.7	Houston	October 2014	CLEANING FEE INCLUDED. Studio w
751	$69	0.7	Fort Worth	February 2014	Stay at our contemporary guesthou
764	$109	1.0	The Woodlands	January 2017	Home sweet home!\n\nNewly furn
804	$90	1.0	San Antonio	July 2014	Our newest available space on the

which means all the "studio" have been changed to 0.7 instead.

3.2.2 Transformation of Dollar Symbol

When trying to calculate the sum and mean of the "average_rate_per_night" column, we found that this column is not in numeric format and cause a lot of trouble. Therefore, it is vital to get rid of the "$" sign and change the data to numeric. Here is the screen shot before change:

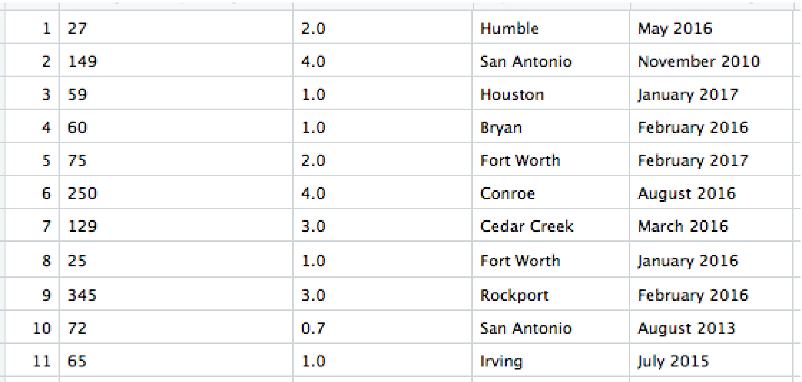


X	average_rate_per_night	bedrooms_count	city	date_of_listing
1	$27	2.0	Humble	May 2016
2	$149	4.0	San Antonio	November 2010
3	$59	1.0	Houston	January 2017
4	$60	1.0	Bryan	February 2016
5	$75	2.0	Fort Worth	February 2017
6	$250	4.0	Conroe	August 2016
7	$129	3.0	Cedar Creek	March 2016
8	$25	1.0	Fort Worth	January 2016
9	$345	3.0	Rockport	February 2016
10	$72	0.7	San Antonio	August 2013
11	$65	1.0	Irving	July 2015

and we can see the ”$” symbol in the table. Now use the following R code:

```
# Convert all the entries from dollars to numeric format
airbnb_data$average_rate_per_night <- as.numeric(sub('$','',as.character(
  airbnb_data$average_rate_per_night),fixed=TRUE))
```

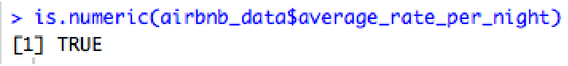
in which way the numbers are transformed as follows:



1	27	2.0	Humble	May 2016
2	149	4.0	San Antonio	November 2010
3	59	1.0	Houston	January 2017
4	60	1.0	Bryan	February 2016
5	75	2.0	Fort Worth	February 2017
6	250	4.0	Conroe	August 2016
7	129	3.0	Cedar Creek	March 2016
8	25	1.0	Fort Worth	January 2016
9	345	3.0	Rockport	February 2016
10	72	0.7	San Antonio	August 2013
11	65	1.0	Irving	July 2015

To check if the data is in the numeric format, use the code ”is.numeric” and here is the result:

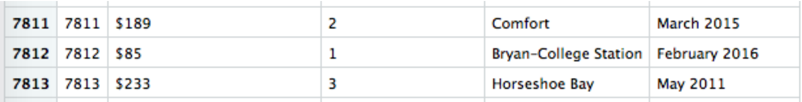
```
is.numeric(airbnb_data$average_rate_per_night)
```



```
> is.numeric(airbnb_data$average_rate_per_night)
[1] TRUE
```

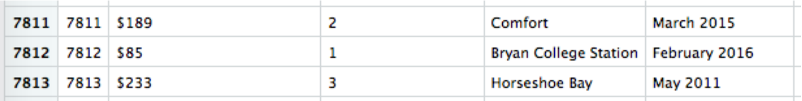
3.2.3 Transformation of Unformatted Symbol

When we had an overview of the dataset, we found that the symbol ”” cannot be read, especially on a Mac. Therefore, it is suggested to get rid of the ”” sign in the entry of ”BryanCollege Station”. Here are the screen shots of beforeimage, R code and after-image:



7811	7811	$189	2	Comfort	March 2015
7812	7812	$85	1	Bryan-College Station	February 2016
7813	7813	$233	3	Horseshoe Bay	May 2011

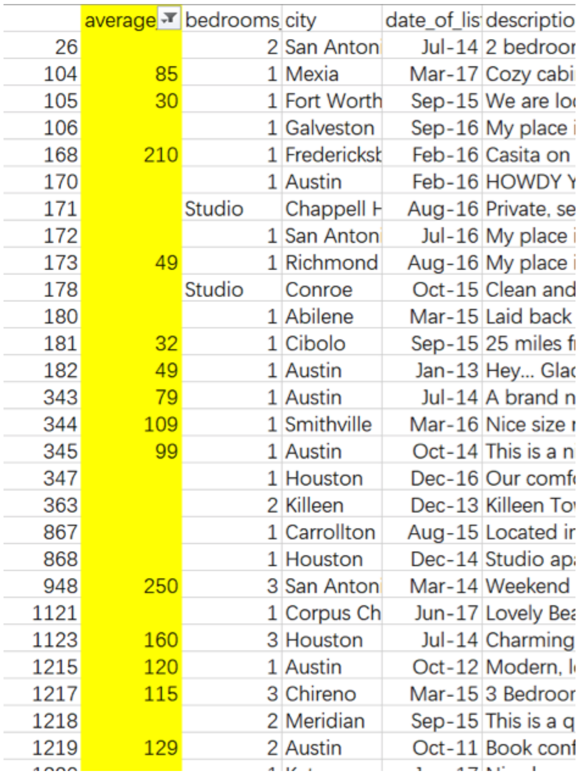
```
airbnb_data$city[which(airbnb_data$city == "Bryan-College Station")] = "Bryan College Station"
```



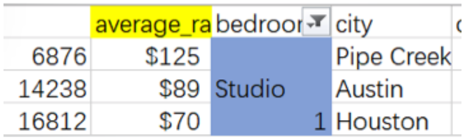
7811	7811	$189	2	Comfort	March 2015
7812	7812	$85	1	Bryan College Station	February 2016
7813	7813	$233	3	Horseshoe Bay	May 2011

3.3 Missing Data

As we know that some prices are missing and some properties do not have the number of bedrooms, it is necessary to fix them. We have checked in chapter 2.3 that there are about 20 of them. As this is not a large number, we firstly check with their website, shown in the ”URL” column. This method helps find out 14 prices of them and 2 for the number of bedrooms, which can then be inserted into the dataset. Here is the Excel collection of data got from the Internet:

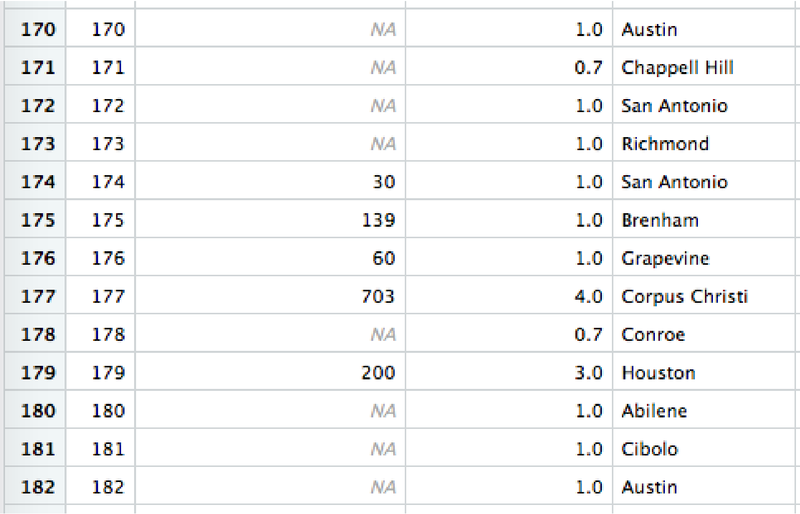


	average	bedrooms	city	date_of_lis	descriptio
26		2	San Anton	Jul-14	2 bedroor
104	85	1	Mexia	Mar-17	Cozy cabi
105	30	1	Fort Worth	Sep-15	We are lo
106		1	Galveston	Sep-16	My place i
168	210	1	Fredericksb	Feb-16	Casita on
170		1	Austin	Feb-16	HOWDY Y
171		Studio	Chappell H	Aug-16	Private, se
172		1	San Anton	Jul-16	My place i
173	49	1	Richmond	Aug-16	My place i
178		Studio	Conroe	Oct-15	Clean and
180		1	Abilene	Mar-15	Laid back
181	32	1	Cibolo	Sep-15	25 miles f
182	49	1	Austin	Jan-13	Hey... Glac
343	79	1	Austin	Jul-14	A brand n
344	109	1	Smithville	Mar-16	Nice size r
345	99	1	Austin	Oct-14	This is a n
347		1	Houston	Dec-16	Our comf
363		2	Killeen	Dec-13	Killeen To
867		1	Carrollton	Aug-15	Located ir
868		1	Houston	Dec-14	Studio ap
948	250	3	San Anton	Mar-14	Weekend
1121		1	Corpus Ch	Jun-17	Lovely Bea
1123	160	3	Houston	Jul-14	Charming
1215	120	1	Austin	Oct-12	Modern, l
1217	115	3	Chireno	Mar-15	3 Bedroor
1218		2	Meridian	Sep-15	This is a q
1219	129	2	Austin	Oct-11	Book conf



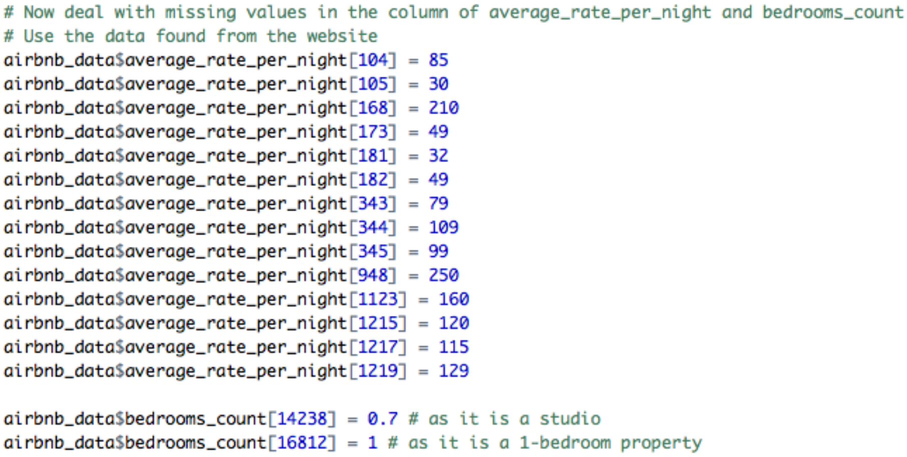
	average_ra	bedroor	city
6876	$125		Pipe Creek
14238	$89	Studio	Austin
16812	$70	1	Houston

For the convenience of comparison, here is a screen shot from row 170 which has the most frequent missing data:



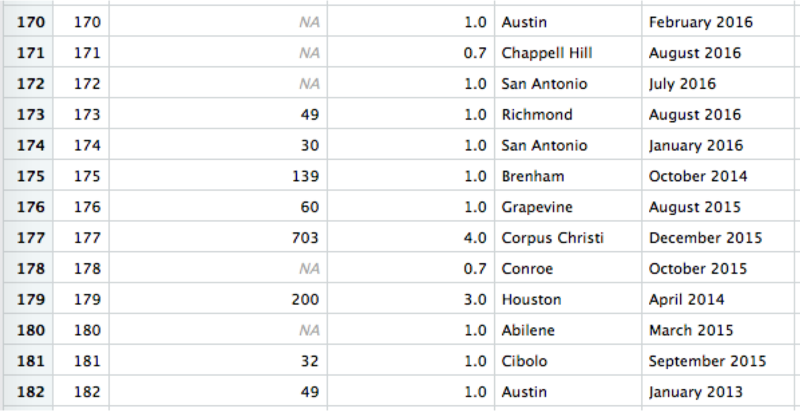
170	170	NA	1.0	Austin
171	171	NA	0.7	Chappell Hill
172	172	NA	1.0	San Antonio
173	173	NA	1.0	Richmond
174	174	30	1.0	San Antonio
175	175	139	1.0	Brenham
176	176	60	1.0	Grapevine
177	177	703	4.0	Corpus Christi
178	178	NA	0.7	Conroe
179	179	200	3.0	Houston
180	180	NA	1.0	Abilene
181	181	NA	1.0	Cibolo
182	182	NA	1.0	Austin

By using the following code, we insert all the values we found into the dataset:



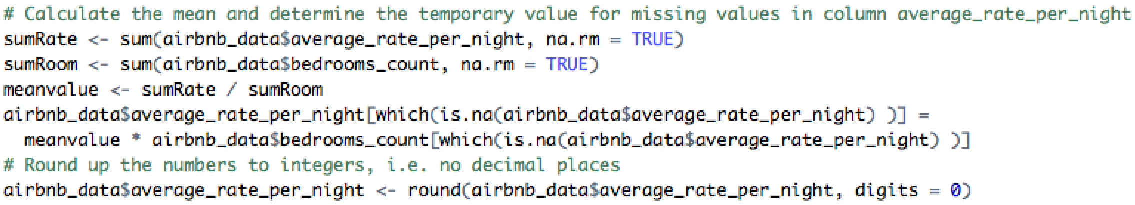
```
# Now deal with missing values in the column of average_rate_per_night and bedrooms_count
# Use the data found from the website
airbnb_data$average_rate_per_night[104] = 85
airbnb_data$average_rate_per_night[105] = 30
airbnb_data$average_rate_per_night[168] = 210
airbnb_data$average_rate_per_night[173] = 49
airbnb_data$average_rate_per_night[181] = 32
airbnb_data$average_rate_per_night[182] = 49
airbnb_data$average_rate_per_night[343] = 79
airbnb_data$average_rate_per_night[344] = 109
airbnb_data$average_rate_per_night[345] = 99
airbnb_data$average_rate_per_night[948] = 250
airbnb_data$average_rate_per_night[1123] = 160
airbnb_data$average_rate_per_night[1215] = 120
airbnb_data$average_rate_per_night[1217] = 115
airbnb_data$average_rate_per_night[1219] = 129

airbnb_data$bedrooms_count[14238] = 0.7 # as it is a studio
airbnb_data$bedrooms_count[16812] = 1 # as it is a 1-bedroom property
```

and here is the result:



170	170	NA	1.0	Austin	February 2016
171	171	NA	0.7	Chappell Hill	August 2016
172	172	NA	1.0	San Antonio	July 2016
173	173	49	1.0	Richmond	August 2016
174	174	30	1.0	San Antonio	January 2016
175	175	139	1.0	Brenham	October 2014
176	176	60	1.0	Grapevine	August 2015
177	177	703	4.0	Corpus Christi	December 2015
178	178	NA	0.7	Conroe	October 2015
179	179	200	3.0	Houston	April 2014
180	180	NA	1.0	Abilene	March 2015
181	181	32	1.0	Cibolo	September 2015
182	182	49	1.0	Austin	January 2013

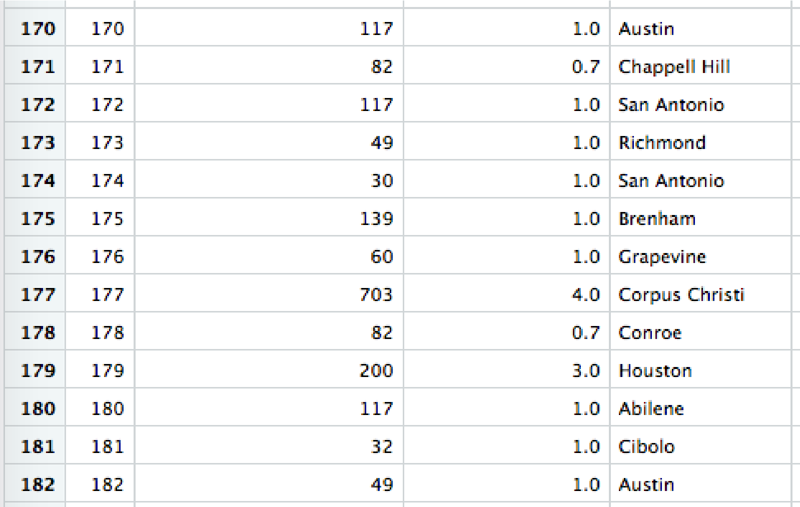
We can see that there are still a number of entries missing. We would like to calculate a mean and use the mean for the substitution. By the code below, we can calculate the average price per room, and then multiply by the number of rooms accordingly, to get the substitution for each missing entry. The code for calculating the mean is:



```
# Calculate the mean and determine the temporary value for missing values in column average_rate_per_night
sumRate <- sum(airbnb_data$average_rate_per_night, na.rm = TRUE)
sumRoom <- sum(airbnb_data$bedrooms_count, na.rm = TRUE)
meanvalue <- sumRate / sumRoom
airbnb_data$average_rate_per_night[which(is.na(airbnb_data$average_rate_per_night) )] =
  meanvalue * airbnb_data$bedrooms_count[which(is.na(airbnb_data$average_rate_per_night) )]
# Round up the numbers to integers, i.e. no decimal places
airbnb_data$average_rate_per_night <- round(airbnb_data$average_rate_per_night, digits = 0)
```

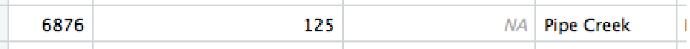
The mean price is calculated to be $116.9 per night per room.

As a result, all the entries in the ”average_rate_per_night” are fixed:

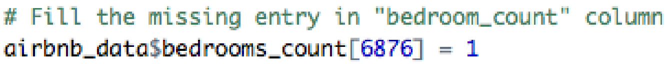


170	170	117	1.0	Austin
171	171	82	0.7	Chappell Hill
172	172	117	1.0	San Antonio
173	173	49	1.0	Richmond
174	174	30	1.0	San Antonio
175	175	139	1.0	Brenham
176	176	60	1.0	Grapevine
177	177	703	4.0	Corpus Christi
178	178	82	0.7	Conroe
179	179	200	3.0	Houston
180	180	117	1.0	Abilene
181	181	32	1.0	Cibolo
182	182	49	1.0	Austin

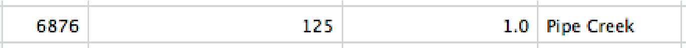
The only left entry in the ”bedroom_count” would be filled with 1-bedroom property:



6876	125	*NA*	Pipe Creek



```
# Fill the missing entry in "bedroom_count" column
airbnb_data$bedrooms_count[6876] = 1
```



6876	125	1.0	Pipe Creek

In this way, all the problems in the dataset raised before are fixed.

Chapter 4

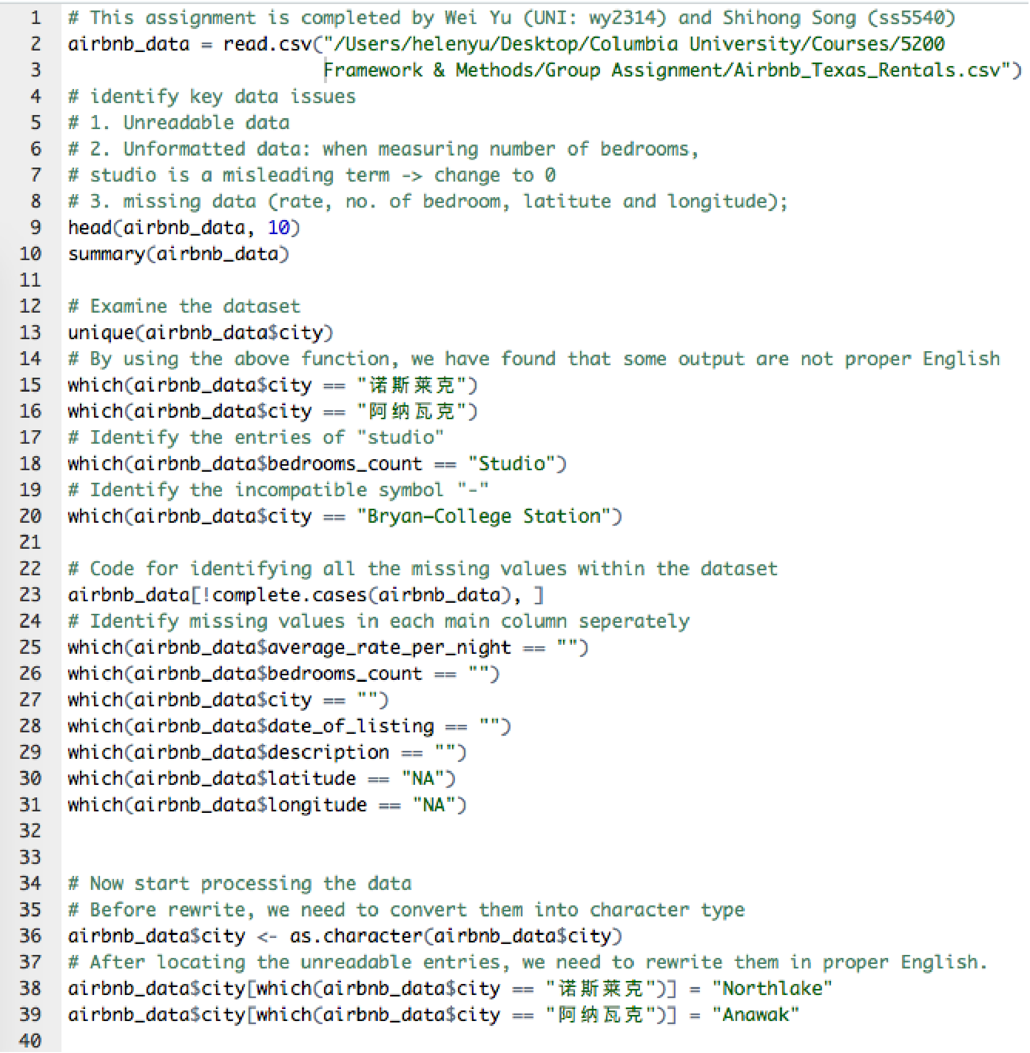
Conclusion

In this assignment, we downloaded a dataset from the Kaggles, which is about Airbnb properties in Texas. The dataset is also referenced from another website, KD nuggets.

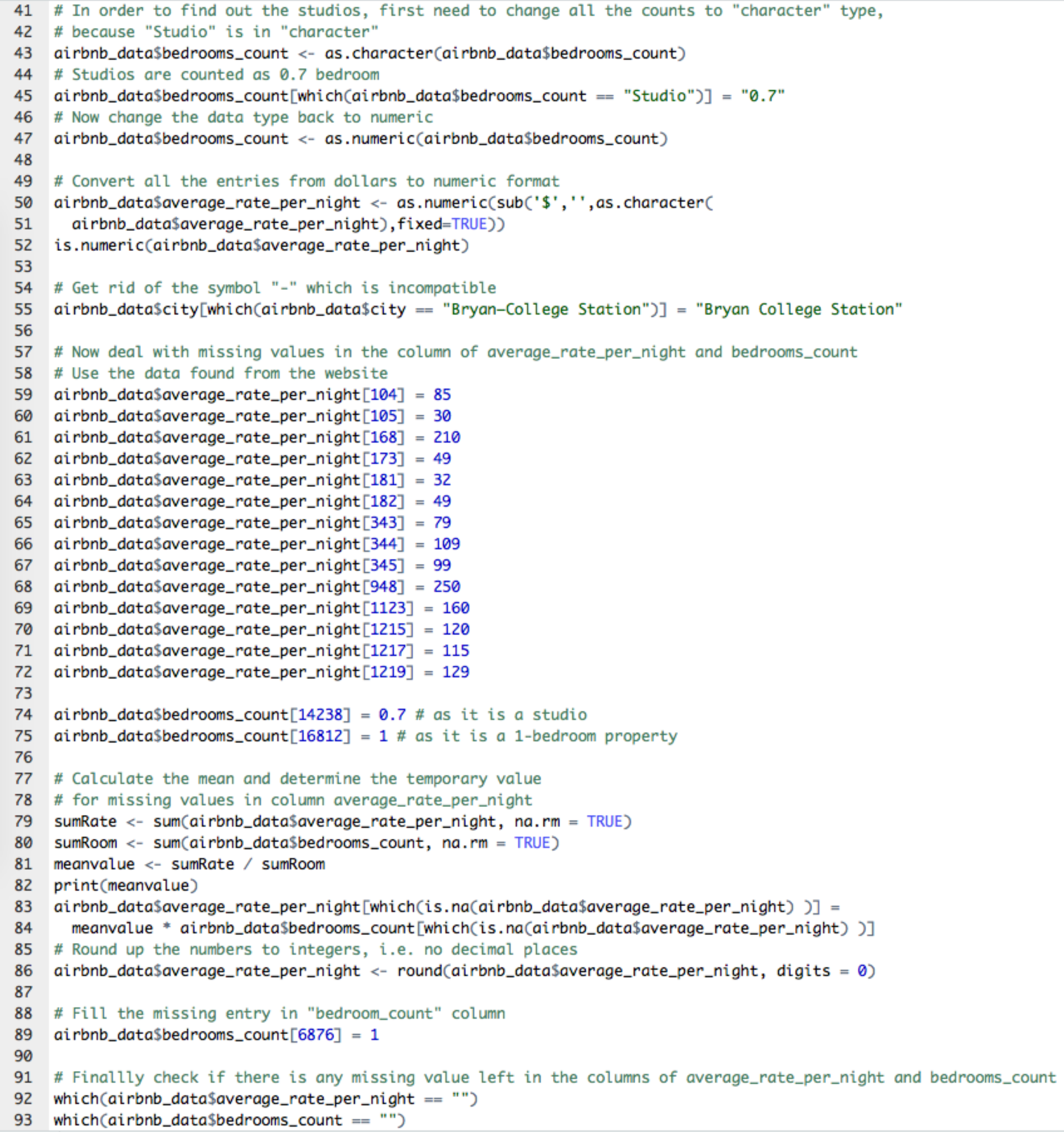
There are some data problems in our dataset, some are important whereas the others can be neglected. We find some problems like unreadable code, missing data and unformatted data and used different tools and codes to fix them in order to get a more comprehensive dataset. Then we describe our solutions to clean up our dataset in chapter 3, using the knowledge we learn from class and website. Finally, we processed our new corrected data and conducted the final analytical dataset.

Appendices

R Code Screen Shots



```
# This assignment is completed by Wei Yu (UNI: wy2314) and Shihong Song (ss5540)
airbnb_data = read.csv("/Users/helenyu/Desktop/Columbia University/Courses/5200
                       Framework & Methods/Group Assignment/Airbnb_Texas_Rentals.csv")
# identify key data issues
# 1. Unreadable data
# 2. Unformatted data: when measuring number of bedrooms,
# studio is a misleading term -> change to 0
# 3. missing data (rate, no. of bedroom, latitute and longitude);
head(airbnb_data, 10)
summary(airbnb_data)

# Examine the dataset
unique(airbnb_data$city)
# By using the above function, we have found that some output are not proper English
which(airbnb_data$city == "诺斯莱克")
which(airbnb_data$city == "阿纳瓦克")
# Identify the entries of "studio"
which(airbnb_data$bedrooms_count == "Studio")
# Identify the incompatible symbol "-"
which(airbnb_data$city == "Bryan-College Station")

# Code for identifying all the missing values within the dataset
airbnb_data[!complete.cases(airbnb_data), ]
# Identify missing values in each main column seperately
which(airbnb_data$average_rate_per_night == "")
which(airbnb_data$bedrooms_count == "")
which(airbnb_data$city == "")
which(airbnb_data$date_of_listing == "")
which(airbnb_data$description == "")
which(airbnb_data$latitude == "NA")
which(airbnb_data$longitude == "NA")


# Now start processing the data
# Before rewrite, we need to convert them into character type
airbnb_data$city <- as.character(airbnb_data$city)
# After locating the unreadable entries, we need to rewrite them in proper English.
airbnb_data$city[which(airbnb_data$city == "诺斯莱克")] = "Northlake"
airbnb_data$city[which(airbnb_data$city == "阿纳瓦克")] = "Anawak"

```



```r
# In order to find out the studios, first need to change all the counts to "character" type,
# because "Studio" is in "character"
airbnb_data$bedrooms_count <- as.character(airbnb_data$bedrooms_count)
# Studios are counted as 0.7 bedroom
airbnb_data$bedrooms_count[which(airbnb_data$bedrooms_count == "Studio")] = "0.7"
# Now change the data type back to numeric
airbnb_data$bedrooms_count <- as.numeric(airbnb_data$bedrooms_count)

# Convert all the entries from dollars to numeric format
airbnb_data$average_rate_per_night <- as.numeric(sub('$','',as.character(
  airbnb_data$average_rate_per_night),fixed=TRUE))
is.numeric(airbnb_data$average_rate_per_night)

# Get rid of the symbol "-" which is incompatible
airbnb_data$city[which(airbnb_data$city == "Bryan-College Station")] = "Bryan College Station"

# Now deal with missing values in the column of average_rate_per_night and bedrooms_count
# Use the data found from the website
airbnb_data$average_rate_per_night[104] = 85
airbnb_data$average_rate_per_night[105] = 30
airbnb_data$average_rate_per_night[168] = 210
airbnb_data$average_rate_per_night[173] = 49
airbnb_data$average_rate_per_night[181] = 32
airbnb_data$average_rate_per_night[182] = 49
airbnb_data$average_rate_per_night[343] = 79
airbnb_data$average_rate_per_night[344] = 109
airbnb_data$average_rate_per_night[345] = 99
airbnb_data$average_rate_per_night[948] = 250
airbnb_data$average_rate_per_night[1123] = 160
airbnb_data$average_rate_per_night[1215] = 120
airbnb_data$average_rate_per_night[1217] = 115
airbnb_data$average_rate_per_night[1219] = 129

airbnb_data$bedrooms_count[14238] = 0.7 # as it is a studio
airbnb_data$bedrooms_count[16812] = 1 # as it is a 1-bedroom property

# Calculate the mean and determine the temporary value
# for missing values in column average_rate_per_night
sumRate <- sum(airbnb_data$average_rate_per_night, na.rm = TRUE)
sumRoom <- sum(airbnb_data$bedrooms_count, na.rm = TRUE)
meanvalue <- sumRate / sumRoom
print(meanvalue)
airbnb_data$average_rate_per_night[which(is.na(airbnb_data$average_rate_per_night) )] =
  meanvalue * airbnb_data$bedrooms_count[which(is.na(airbnb_data$average_rate_per_night) )]
# Round up the numbers to integers, i.e. no decimal places
airbnb_data$average_rate_per_night <- round(airbnb_data$average_rate_per_night, digits = 0)

# Fill the missing entry in "bedroom_count" column
airbnb_data$bedrooms_count[6876] = 1

# Finallly check if there is any missing value left in the columns of average_rate_per_night and bedrooms_count
which(airbnb_data$average_rate_per_night == "")
which(airbnb_data$bedrooms_count == "")
```

Bibliography

[1] Inc. Airbnb. About us - airbnb newsroom. https://press.atairbnb.com/about-us/, 2018.

[2] Wikipedia contributors. Airbnb — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Airbnb&oldid=828363115, 2018. [Online; accessed 2-March-2018].

[3] etc. Donyoe, Faraz92. Airbnb property data from texas - kaggle. https://www.kaggle.com/PromptCloudHQ/airbnb-property-data-from-texas, 2018.