Cleaning and Processing Dataset of Airbnb Property from Texas

APAN 5200 PROJECT DELIVERABLE
BY
WEI YU

BACKGROUND

Airbnb:

- Founded in 2008, online platform for landlord to rent their houses and/or rooms to visitors who are traveling to those cities.
- Over 4 millions listing in 65,000 cities and 191 countries

Kaggle Dataset:

- Size: more than 18,000 property listings from Texas, US.
- Headings: Average Rate Per Night, Bedrooms Count, City, Date of Listing, Description, Latitude, Longitude, Title, Property description, and URL.

Research Question:

- 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 various cities in Texas?
- 3.What is the trend of the listings on Airbnb each year?

PROBLEM 1 – UNFORMATTED CODE

Price - not in number format - cannot do calculations

```
> mean(airbnb_data$average_rate_per_night)
[1] NA
Warning message:
In mean.default(airbnb_data$average_rate_per_night) :
  argument is not numeric or logical: returning NA
```

Convert prices to numeric:

```
# 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))
```

Check:

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

Text Format of Average Rate per Night:

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

Numeric Format of Average Rate per Night:

1	27	2.0	Humble	May 2016
2	149	4.0	San Antonio	November 2
3	59	1.0	Houston	January 201
4	60	1.0	Bryan	February 20
5	75	2.0	Fort Worth	February 20
6	250	4.0	Conroe	August 201
7	129	3.0	Cedar Creek	March 2016
8	25	1.0	Fort Worth	January 201
9	345	3.0	Rockport	February 20
10	72	0.7	San Antonio	August 201

PROBLEM 2 – IMPROPER DATA TYPE

Lots of “Studios” in the list of “bedrooms_count”:

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
125	$55	Studio	Houston	May 2017	The studio apartm
134	$98	Studio	College Station	June 2016	A uniquely styled t
157	$50	Studio	Cleburne	December 2016	My place is wonder

Transform “Studios” to count numerically:

```
# 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)
```

“Studios” are counted to be 0.7 bedrooms:

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 cultu
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

PROBLEM 3 – MISSING DATA

Find the missing data:

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

	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
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
173	49	1	Richmond	Aug-16	My place
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
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

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

Find the available data from the web:

J
url
https://www.airbnb.com/rooms/18520444?location=Cleveland%2C%20TX
https://www.airbnb.com/rooms/17481455?location=Cibolo%2C%20TX
https://www.airbnb.com/rooms/16926307?location=Beach%20City%2C%20TX
https://www.airbnb.com/rooms/11839729?location=College%20Station%2C%20TX
https://www.airbnb.com/rooms/17325114?location=Colleyville%2C%20TX

Input the new data:

```
# 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
```

PROBLEM 3 – MISSING DATA

Find Undefined Missing data:

170	170	NA	1.0	Austin
171	171	NA	0.7	Chappell Hill
172	172	NA	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	NA	0.7	Conroe
179	179	200	3.0	Houston
180	180	NA	1.0	Abilene

Input the mean value:

```
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)
```

X	average_rate_per_night	bedrooms_count	city	date_of_listing
26	234	2.0	San Antonio	July 2014

170	170	117	1.0	Austin
171	171	[illegible]	0.7	Chappell Hill
172	172	117	1.0	San Antonio
173	173	[illegible]	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	[illegible]	3.0	Houston
180	180	117	1.0	Abilene

Estimate the mean value:

```
# Calculate the mean and determine the temporary value for missing values in
sumRate <- sum(airbnb_data$average_rate_per_night, na.rm = TRUE)
sumRoom <- sum(airbnb_data$bedrooms_count, na.rm = TRUE)
meanvalue <- sumRate / sumRoom
```

CONCLUSION

Error Types: Unformatted data, Improper data type, Missing data

How did we solve:

- Format in the right way
- Substitute with proper data type
- Find the right ones for the missings or calculate alternatives

What to improve:

- A few columns not considered - title and descriptions - in text
- Check keywords in title and descriptions - see the relevance to price or ratings
- Check longitudes and latitudes - draw a density map for the locations of properties