Price of NYC Airbnb Listings

November 9, 2022

Names	Student Numbers	Contributions
Xingyu Xian(group leader)	46865887	Writing
Yifan Lu	24759672	Data analysis
Yiran Liu	10751485	Part II
Zhengling Jiang	82219353	R code
Wenjing Guo	95008447	Writing

PART I

1 Introduction

1.1 Background

Airbnb is a digital marketing firm that links people looking for lodging with those looking to rent out their properties for a short-term or long-term period time. It connects people who want to rent with those looking for accommodation. Apartments are the most common type of rental property, but there are also residences, boats, and many other types. Airbnb has expanded the traveling possibilities, and hosts and guests are creating a new and extraordinary way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

1.2 Objectives

To investigate the average price and the proportion of listings over 150 USD of Airbnb in New York City for 2019.

1.3 Significance

As more people are choosing Airbnb as a medium to select accommodation, the need to investigate the average price has arisen for travelers. New York, being one of the most prosperous cities in the world, is an appropriate place to research. New York City has a highly active Airbnb market with more than 48000 listings as of August 2019. Investigating the average price of Airbnb in New York City for 2019 is helpful as a reference for budgeting for tourists and for hosts.

Secondly, the proportion of listing prices over 150 USD will be investigated. According to a Zippia research, the personal income in the U.S. is around 63,214 USD, and 5% of it is spent on travel.[1] According to an Expedia survey, Americans travel domestically about twice a year and spend 10% on accommodation every day, so the assumption made here is that 150 USD is the daily accommodation budget of most people.[2] By

[1] Zippia. "25+ Essential Average American Income Statistics [2022]: Household + Personal Income In The US" Zippia.com. Oct. 26, 2022, https://www.zippia.com/advice/average-american-income/

[2] Eveline. “How Often Does the Average American Travel?” Journeyz, 11 May 2022, https://journeyz.co/often-average-american-travel/

calculating the proportion of listing prices over 150 USD, we can estimate how many listings are affordable to most people.

2 Data Selection and Summaries

2.1 Population and Parameter of Interest

Since the objectives are investigating the average price and the proportion of listings over 150 USD of Airbnb in New York City for 2019, the targeted population is all listings on Airbnb in New York City for 2019. We have two parameters of interest. The first parameter is the average price of Airbnb listings in New York in 2019, whereas the second parameter is the proportion of Airbnb listings exceeding 150 USD.

2.2 Data selection

Our dataset is obtained from Kaggle, which is an open data source website. The original source can be found on Inside Airbnb. This dataset describes the listing activity and metrics in NYC, NY for 2019. There are 16 variables in the dataset. The continuous variables are listing id, host id, latitude, longitude, price, minimum nights, the number of reviews, reviews per month, calculated host listings count, and availability 365. The categorical variables are name, hostname, neighborhood group, neighborhood, room type, and last review. Since we are analyzing the prices of the listings and the proportion of listings exceeding 150 USD, we will be using price and room type.

2.3 Sampling Methods

The methods going to be used are Simple Random Sampling and stratified random sampling.

A simple random sample is a randomly selected subset of a population. In this sampling method, each member of the population has an exactly equal chance of being selected. This method is the most straightforward of all the probability sampling methods since it only involves a single random selection and requires little advanced knowledge about the population. Because it uses randomization, any research performed on this sample

should have high internal and external validity. Simple random sampling works best as we are studying a limited population that can easily be sampled.

Stratified sampling is appropriate when we want to ensure that specific characteristics are proportionally represented in the sample. Stratified sampling refers to partitioning the population into H mutually exclusive and exhaustive subsets or strata. In each stratum, an SRS is obtained. The results from all strata are then combined to estimate the overall population characteristic. We can synthesize separate samples from sub-populations (strata) to infer whole-population quantities.

In our sample, the room type variable is chosen to be stratified. Since the goal is to investigate the average price of Airbnb in New York City for 2019, we believe room type may significantly impact the outcome variable. In general, the larger the type of room, the higher the price. For example, the price of the entire room is always higher than the shared room price and private room prices. Then we can have 3 distinct strata, private room, entire home/apt, and shared room.

2.4 Sample Size

In order to calculate the sample size, the following variables are needed: Population size (48895), a margin of error (0.03), Z-score (1.96), and sample population (0.5). We are using p=0.5 which gives a "worst case scenario". The technique of computing sample size is shown in the formula below, and the sample size we obtained is 1045.

$$\frac{\frac{1.96^2 \times 0.5(1-0.5)}{0.03^2}}{1+\frac{1.96^2 \times 0.5(1-0.5)}{0.03^2 \times 48895}}$$

It is not essential to use the FPC because the sample size (1045) / population size (48895) is less than 0.05. However, we decided to apply FPC to improve the precision of the output.

The numbers for the three different types of rooms were as follows: entire home/apartment (25409), private room (22326), and shared room (1160). Given the population size is 48895, the proportion of each of the

room types is as follows: entire home/apartment (0.52), private room (0.46), and shared room (0.024). Therefore, through stratified sampling with proportional allocation:
the sample size for the entire home/apartment is 1045 x 0.52 = 543;
the sample size for a private room is 1045 x 0.4566 = 477;
the sample size for a shared room is 1045 x 0.0237 = 25.

3 Data analysis

3.1 The estimated average price of listings

For SRS, the estimated average price of listings is 165.44 USD with SE of 11.91, and a 95 percent confidence interval of (142.35, 188.53). For stratified sampling, the estimated average price of listings is 149.83 USD with an SE of 4.84, and a 95 percent confidence interval of (140.44, 159.22). The SE for stratified (4.84) is significantly lower than the SE for SRS (11.91). Hence the SRS method is preferred.

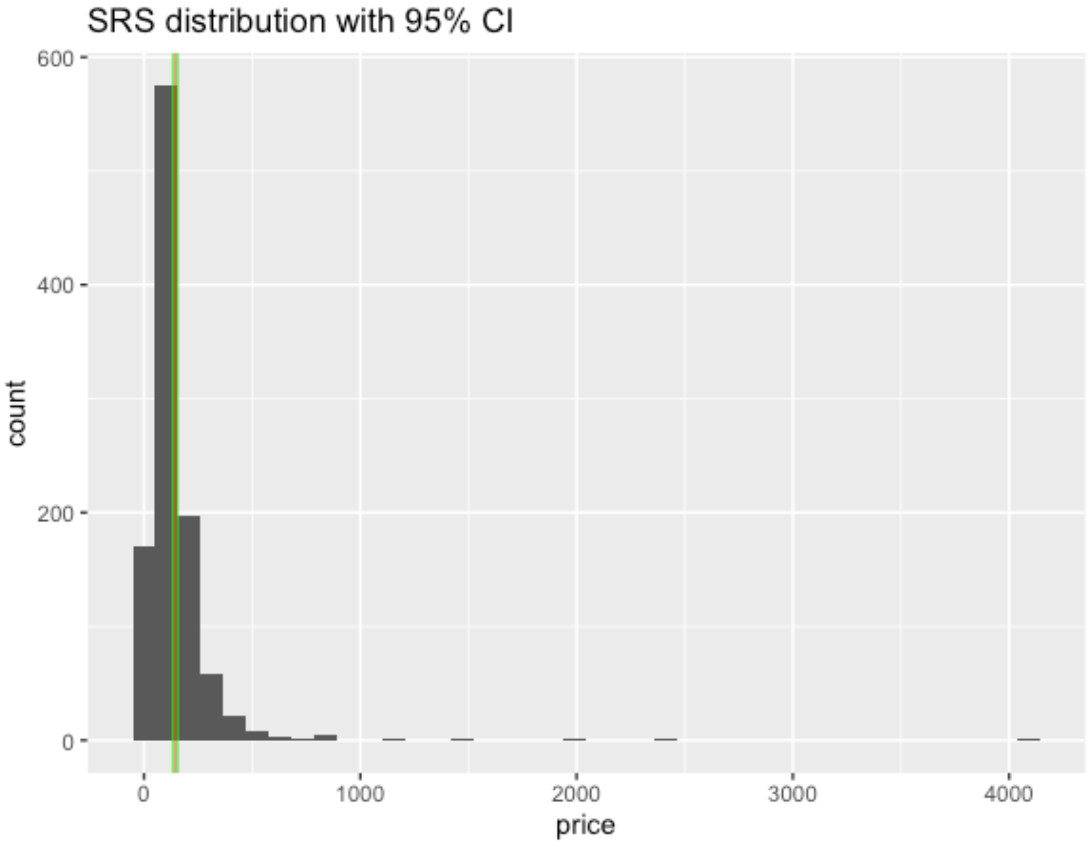

Figure 1: SRS distribution with 95% CI. The red line is the estimated mean, and the green lines indicate the confidence interval.

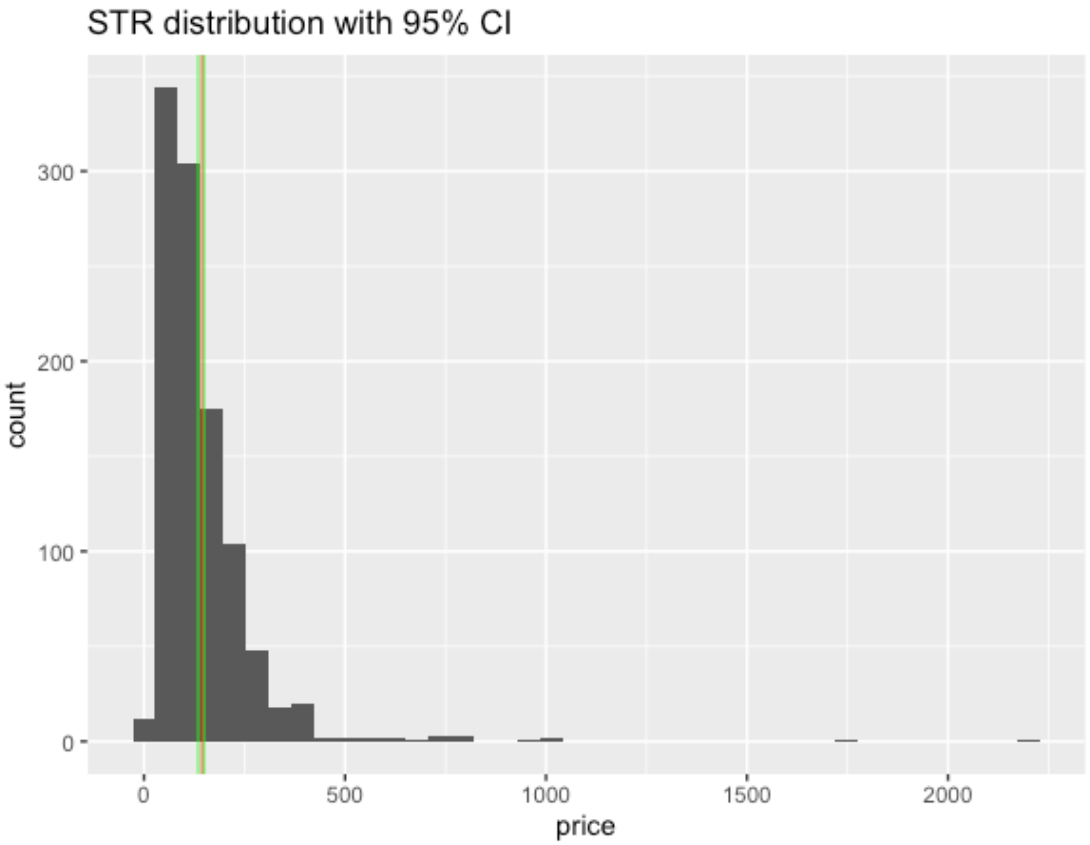

Figure 2: Stratified distribution with 95% CI. The red line is the estimated mean, and the green lines indicate the confidence interval.

3.2 The estimated proportion of listings of price exceeding 150 USD

For SRS, the estimated proportion of listings of price exceeding 150 USD is 0.31 with SE of 0.014, and a 95 percent of confidence interval of (0.29, 0.34). For stratified sampling, the estimated proportion of listings of price exceeding 150 USD is 0.31 with SE of 0.012, and a 95 percent of confidence interval of (0.28, 0.33). The SE for stratified which is 0.012 is lower than the SE for SRS which is 0.014.

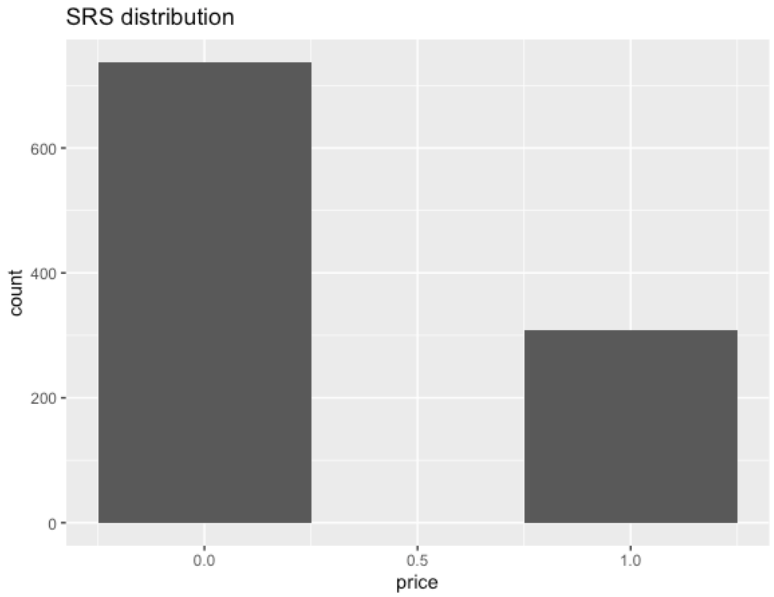

Figure 3: SRS distribution of 0.0 showing the number of listings below and equal to 150 USD and 1.0 showing the number of listings over 150 USD

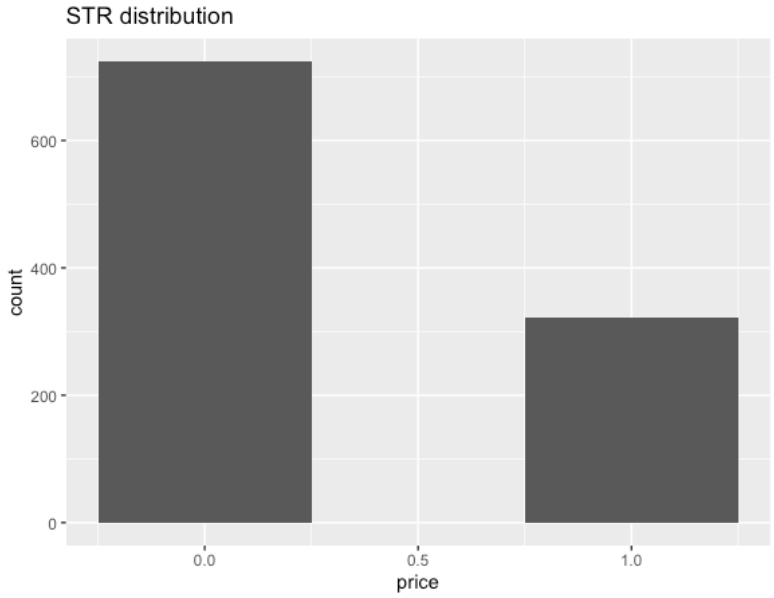

Figure 4: Stratified distribution of 0.0 showing the number of listings below and equal to 150 USD and 1.0 showing the number of listings over 150 USD

Simple random sampling may be easy to carry out and only requires a basic understanding of the population of listings through Airbnb in New York City for 2019. It also results in less bias. However, since we stratify the entire population of listings through Airbnb in New York City for 2019 before utilizing random sample approaches, stratified sampling can ensure that the samples we take have an accurate representation of each of the three room types. Specifically, stratified random sampling provides better

coverage of the population since we can control the various room types to ensure that they are all represented in the sampling through proportional allocation. In addition, the SE of the stratified sampling approach is lower when compared to the standard deviation of the SRS. Therefore, the stratified sampling strategy is preferred.

4 Conclusion

4.1 Overall conclusion

Our final conclusions are that the estimated average price is 149.83 USD with an SE of 4.84 and the 95% confidence interval is (140.44, 159.22), and the estimated proportion of listings over 150 USD is 0.31 with SE of 0.012 and the 95% confidence interval is (0.28, 0.33). Thus, we recommend that visitors who want to stay at Airbnb in New York consider 149.83 USD as their rent estimate. Since the proportion of listings over 150 USD is almost 30 percent, there are not a lot of such houses, and tourists can consider more houses under 150 USD.

4.2 Limitation

For data collection, the population size is large and the sample size is relatively small, which may reduce the power of the study. Furthermore, while the price may fluctuate over a year, the dataset we use has only one price for each listing. For example, the price may increase during the tourist season and may decrease during the slack season. The price for each Airbnb should be the average price of the year instead of the instantaneous value. Thus, the dataset may cause inaccuracy when we estimate the average price of Airbnb in NYC in 2019. In addition, Since the dataset is collected in 2019, the result may not hold for other years.

Our sample size depends on population size, we cannot guarantee that our results are still reasonable if the population size becomes large. Furthermore, we stratified with proportion allocation, the proportion for different room types may not hold if the population size and the weighted averages have changed.

5 Appendix

5.1 Dataset source

Dgomonov. (2019, August 12). *New York City airbnb open data*. Kaggle. Retrieved November 7, 2022,
from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

5.2 Dataset overview

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room
2	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt
3	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room
4	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt
5	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt

price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
149	1	9	2018-10-19	0.21	6	365
225	1	45	2019-05-21	0.38	2	355
150	3	0		NA	1	365
89	1	270	2019-07-05	4.64	1	194
80	10	9	2018-11-19	0.10	1	0

5.3 Codes

Continuous parameter

```
NYC <- read.csv("NYC.csv")

attach(NYC)
# population size for different room types
N.h <- tapply(price, room_type, length)
# name of the room_types
room_types <- names(N.h)
N <- sum(N.h)
true.value <- mean(NYC$price)
true.value

N.h

# SRS
set.seed(1234)
# sample size
n <- 1045
SRS.indices <- sample.int(N,    n, replace = F)
SRS.sample <- NYC[SRS.indices , ]
```

```
ybar.srs <- mean(SRS.sample$price)
se.srs <- sqrt((1 - n / N) * var(SRS.sample$price) / n)
srs <- c(ybar.srs, se.srs)
srs

# 95% confidence interval for SRS
c(ybar.srs - 1.96*sqrt(1 - n / N)*se.srs, ybar.srs + 1.96*sqrt(1 - n
/N)*se.srs)

# STR
# generate a stratified sample
n.h.prop <- round( (N.h/N) * n)
STR.sample.prop <- NULL
set.seed(1234)
for (i in 1: length(room_types))
{
   row.indices <- which(NYC$room_type == room_types[i])
   sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
   STR.sample.prop <- rbind(STR.sample.prop, NYC[sample.indices, ])
}

# STR estimation
ybar.h.prop <- tapply(STR.sample.prop$price,
STR.sample.prop$room_type, mean)
var.h.prop <- tapply(STR.sample.prop$price,
STR.sample.prop$room_type, var)
se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
rbind(ybar.h.prop, se.h.prop)

ybar.str.prop <- sum(N.h / N * ybar.h.prop)
se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
str.prop <- c(ybar.str.prop, se.str.prop)
str.prop

# 95% confidence interval for STR
c(ybar.str.prop - 1.96*sqrt(1 - n / N)*se.str.prop,
   ybar.str.prop + 1.96*sqrt(1 - n / N)*se.str.prop)
```

```
# compare SRS and STR
rbind(srs,    str.prop)
```

Binary parameter

```
# SRS
# We use the sample we get from the continuous parameter part
b.srs.sample <- ifelse(SRS.sample$price > 150, 1, 0)
pbar.srs <- mean(b.srs.sample)
p_se.srs <- sqrt((1 - n / N) * pbar.srs * (1 - pbar.srs) / n)
b.srs <- c(pbar.srs, p_se.srs)
b.srs

# 95% confidence interval
c(pbar.srs - 1.96*sqrt(1 - n / N)*p_se.srs, pbar.srs + 1.96*sqrt(1 - n /
N)*p_se.srs)

# STR
# We use the stratified sample we get from the continuous parameter part
# STR estimation
STR.sample.prop$price <- ifelse(STR.sample.prop$price > 150, 1, 0)
pbar.h.prop <- tapply(STR.sample.prop$price,
STR.sample.prop$room_type, mean)
p_var.h.prop <- pbar.h.prop * (1 - pbar.h.prop)
p_se.h.prop <- sqrt((1 - n.h.prop / N.h) * p_var.h.prop / n.h.prop)
rbind(pbar.h.prop, p_se.h.prop)

pbar.str.prop <- sum(N.h / N * pbar.h.prop)
p_se.str.prop <- sqrt(sum((N.h / N)^2 * p_se.h.prop^2))
b.str.prop <- c(pbar.str.prop, p_se.str.prop)
b.str.prop

# 95% confidence interval
 c(pbar.str.prop - 1.96*sqrt(1 - n / N)*p_se.str.prop, pbar.str.prop +
1.96*sqrt(1 - n / N)*p_se.str.prop)
```

Plot

Continuous parameter:

SRS:

```
srs.plot<-SRS.sample %>% ggplot() + geom_histogram(aes(x = price), bins = 40) + geom_vline(xintercept = ybar.srs, color = "red", alpha=.6, lwd=0.5) + geom_vline(xintercept = srs.ci[1], color = "green", alpha=.6, lwd=0.5) + geom_vline(xintercept = srs.ci[2], color = "green", alpha=.6, lwd=0.5) + ggtitle("SRS distribution with 95% CI")
```

STR:

```
str.plot <- STR.sample.prop %>% ggplot() + geom_histogram(aes(x = price), bins = 40) + geom_vline(xintercept = ybar.srs, color = "red", alpha=.6, lwd=0.5) + geom_vline(xintercept = str.ci[1], color = "green", alpha=.6, lwd=0.5) + geom_vline(xintercept = str.ci[2], color = "green", alpha=.6, lwd=0.5) + ggtitle("STR distribution with 95% CI")
```

Binary parameter:

SRS:

```
b.srs.plot <- b.srs.sample %>% ggplot() + geom_histogram(aes(price), bins = 3) + ggtitle("SRS distribution")
```

STR:

```
b.str.plot <- b.str.sample %>% ggplot() + geom_histogram(aes(price), bins = 3) + ggtitle("STR distribution")
```

5.4 Citations

- Zippia. "25+ Essential Average American Income Statistics [2022]: Household + Personal Income In The US" Zippia.com. Oct. 26, 2022, https://www.zippia.com/advice/average-american-income/
- Eveline. “How Often Does the Average American Travel?” Journeyz, 11 May 2022, https://journeyz.co/often-average-american-travel/.