Zomato Question & Sample test

Module 4: final project milestone 2

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi
Date: Dec 4^{th}, 2022

INTRODUCTION

The data I choose is ZOMATO data. This data is from Kaggle. The data cleaning process was cleaned by referring to Kaggle's python cleaning process (SANSKRUTI, 2022). This data contains numeric data such as prices, ratings, and votes. The dataset is about application named 'ZOMATO' that provides information about restaurants. A big advantage of this dataset is having a lot of categorical data. I will conduct two-sample t-test in this module to figure out what is the answer to my question.

DATASET EXPLANATION & ORIGINAL QUESTION

	A	B	C	D	E	F	G	H	I	J
1	url	address	name	online_order	book_table	rate	votes	phone	location	rest_type
2	https://wv	942, 21st M	Jalsa	Yes	Yes	4.1/5	775	080	Banashankari	Casual Dining
3	https://wv	2nd Floor,	Spice Elepl	Yes	No	4.1/5	787	080 41714	Banashankari	Casual Dining
4	https://wv	1112, Next	San Churrc	Yes	No	3.8/5	918	+91 96634	Banashankari	Cafe, Casual D
5	https://wv	1st Floor, /	Addhuri Uc	No	No	3.7/5	88	+91 96200	Banashankari	Quick Bites

K	L	M	N	O	P	Q
dish_liked	cuisines	approx_cost(f	reviews_list	menu_item	listed_in(type)	listed_in(city)
Pasta, Lunch Buffet, Ma	North Indian,	800	[('Rated 4.0', 'F	[]	Buffet	Banashankari
Momos, Lunch Buffet, C	Chinese, Nort	800	[('Rated 4.0', 'F	[]	Buffet	Banashankari
Churros, Cannelloni, Mi	Cafe, Mexicai	800	[('Rated 3.0', "I	[]	Buffet	Banashankari
Masala Dosa	South Indian,	300	[('Rated 4.0', "I	[]	Buffet	Banashankari

Figure 1 excel file of Zomato dataset

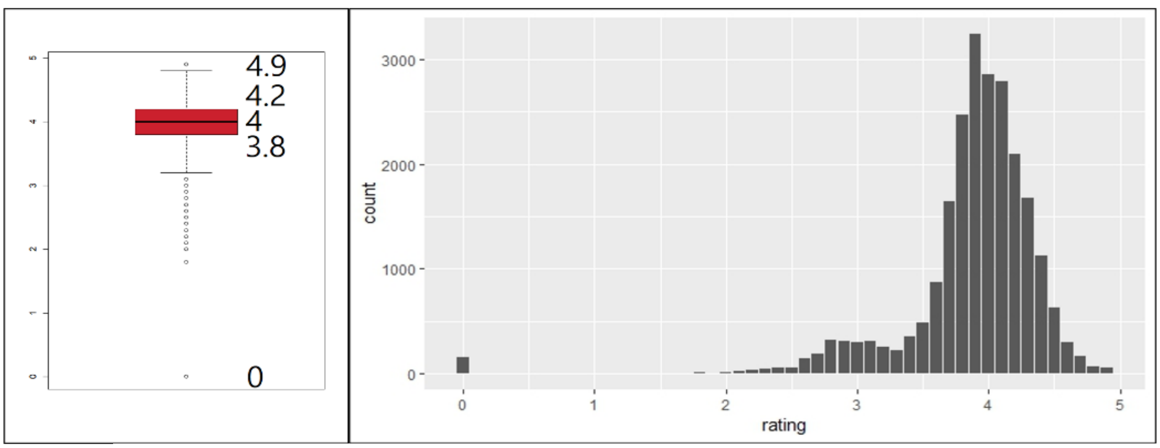

Figure 2 boxplot and histogram of restaurant rating

Explanation

1. This dataset has 23,193 observations of 17 variables.
2. Target data is rating. Numeric data is rating, approximate cost, and number of votes
3. Categorical data is online-order yes or no, book table yes or no, restaurant type, location, city etc.
4. Rating numbers are centralized around 4 but there are also pretty numbers around 3. Upper quartile is 4.2. Median is 4 and lower quartile is 3.8.
5. I contained even 0 rating, because someone who have a terrible experience in restaurant can rate it as 0 sometimes.

Original Question

1. Expensive restaurants have higher rating.
2. If number of voting is high, rating will be higher.
3. The average rating of booking people and not booking people is equal.
4. The average rating of online order and no-online order is equal.
5. The average rating of casual dining and quick bites is same.
6. There is a significant difference between Banashankari and Indiranagar in the average of rating.

Null(H0) and alternative(H1) hypothesis/ Claim=(C)

[1] H0: Average rating between cost over 2000 restaurant and cost below 1000 restaurant is equal.
H1: Average rating of cost over 2000 restaurant is higher than cost below 1000 restaurant (C).

[2] H0: Average rating between votes over 5000 restaurant and votes below 2500 restaurant is equal.
H1: Average rating of votes over 5000 restaurant is higher than votes below 2500 restaurant (C).

[3] H0: Average rating between people who book and who don't book is equal (C).
H1: Average rating between people who book and who don't book is not equal.

[4] H0: Average rating between people who ordered by online and who didn't is equal (C).
H1: Average rating between people who ordered by online and who didn't is not equal.

[5] H0: Average rating between which restaurant type is casual dining and quick bites is equal (C).
H1: Average rating between which restaurant type is casual dining and quick bites is not equal.

[6] H0: Average rating between which location is Banashankari and Indiranagar is equal.
H1: Average rating between which location is Banashankari and Indiranagar is not equal (C).

T-TEST FOR HYPOTHESIS 1

Step 1 Hypothesis.

Rating is higher in cost over 2000 restaurant comparing with below 1000 restaurant.

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

N=30 SAMPLING

```
 Descriptive statistics by group
cost.range: high
            vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
rating         1 30 4.24 0.21    4.3    4.24 0.15 3.7 4.8   1.1 -0.03     0.89 0.04
cost.range*    2 30 1.00 0.00    1.0    1.00 0.00 1.0 1.0   0.0   NaN      NaN 0.00
--------------------------------------------------------------------------
cost.range: low
            vars  n mean  sd median trimmed  mad min max range  skew kurtosis   se
rating         1 30 3.61 0.8   3.85    3.75 0.22   0 4.5   4.5 -3.04    10.84 0.15
cost.range*    2 30 2.00 0.0   2.00    2.00 0.00   2 2.0   0.0   NaN      NaN 0.00
```

Figure 3 descriptive analysis of n=30 sampling of high and low cost.range

Step 2 Find the critical value.

- α = 0.05, one-tailed CI is 0.37, Inf

Step 3 Compute the test value.

- t-value is 4.1263 with t-test (alternative="greater").

Step 4 Make the decision.

- Reject the null hypothesis and accept claim. Since 4.1263 > 0.37

Step 5 Summarize the results.

- Average rating of cost over 2000 restaurant is higher than cost below 1000 restaurant.

Interpretation:

1. Before this t-test, I do with the whole population I have. The p-value is so low, so I found the appropriate number of samples. The procedure does not differ greatly from the one used for large samples but is preferable when the number of observations is less than 60 (bmj, n.d.).
2. I do with sample 30 again and the results were same as do with whole population.
3. I could learn that there is preferable number of samples. If I could use the whole population, the number would be so high or so low.
4. However, the result would be not different greatly between the sample and the population.
5. In the real world, I don't have a time or money to conduct whole population.
6. So, I will study what the exact procedure of sample t-test is.

T-TEST FOR HYPOTHESIS 2

Step 1 Hypothesis.

If number of voting is high, rating will be higher.

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

N=30 SAMPLING

```
 Descriptive statistics by group
votes: less
       vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
rating    1 30  3.8 0.83    3.9    3.93 0.3   0 4.8   4.8 -3.21    12.24 0.15
votes*    2 30  1.0 0.00    1.0    1.00 0.0   1 1.0   0.0   NaN      NaN 0.00
------------------------------------------------------------------------------------
votes: many
       vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
rating    1 30 4.49 0.22    4.5     4.5 0.3 4.1 4.9   0.8 -0.09    -0.86 0.04
votes*    2 30 2.00 0.00    2.0     2.0 0.0 2.0 2.0   0.0   NaN      NaN 0.00
```

Figure 4 descriptive analysis of n=30 sampling of less and many votes

Step 2 Find the critical value.

- α = 0.05, one-tailed t-value is 0.23

Step 3 Compute the test value.

- t-value is 4.3979 with t-test (alternative="greater").

Step 4 Make the decision.

- Reject the null hypothesis and accept claim. Since 4.3979 > 0.23

Step 5 Summarize the results.

- If number of voting is high, rating will be higher.

Interpretation:

1. Actually, I did t-test twice because first sampling contains 0.00 rating in votes 'less' group. The results were same. If the results were different I will think about how could I deal with it. However, what I decide in data explanation was 'I contained even 0 rating, because someone who have a terrible experience in restaurant can rate it as 0 sometimes. Therefore, I will reflect two results as one sample was with 0 and the other sample was not with 0.
2. If the votes number is high, rating will be higher than lower votes < 2500.
3. It was much easier to conduct t-test because I made a code for first t-test. I can use that code to do it again.

T-TEST FOR HYPOTHESIS 3

Step 1 Hypothesis.

The average rating of booking people and not booking people is equal.

- H0: μ1 = μ2 (claim) and H1: μ1 ≠ μ2

N=30 SAMPLING

```
 Descriptive statistics by group
booking: no
         vars  n mean  sd median trimmed  mad min max range  skew kurtosis   se
rating      1 30 3.89 0.3    3.9    3.93 0.15 2.9 4.5   1.6 -1.19     2.55 0.05
booking*    2 30 1.00 0.0    1.0    1.00 0.00 1.0 1.0   0.0   NaN      NaN 0.00
------------------------------------------------------------------------
booking: yes
         vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
rating      1 30 4.04 0.36    4.1    4.09 0.22 2.9 4.6   1.7 -1.39     1.87 0.07
booking*    2 30 2.00 0.00    2.0    2.00 0.00 2.0 2.0   0.0   NaN      NaN 0.00
```

Figure 5 descriptive analysis of n=30 sampling of yes or no booking

Step 2 Find the critical value.

- [1] α = 0.05, one-tailed p-value for 29 (smaller one of n-1) is 0.123
- [2] α = 0.05, one-tailed p-value for 29 (smaller one of n-1) is 0.00032
- [3] α = 0.05, one-tailed p-value for 29 (smaller one of n-1) is 0.00002

Step 3 Compute the test value.

- [1] p-value is 0.0123 with t-test (alternative="two-sided").
- [2] p-value is 0.00032 with t-test (alternative="two-sided").
- [3] p-value is 0.00002 with t-test (alternative="two-sided").

Step 4 Make the decision.

- [1] Reject the null hypothesis. Since p-value 0.123 > 0.05 (0.0 in no)
- [2] Accept the null hypothesis. Since p-value 0.00032 < 0.05
- [3] Accept the null hypothesis. Since p-value 0.00002 < 0.05
- [4] 0.005249 [5] 0.0001081 [7] 0.0002151 [9] 0.0002112 [10] 0.00001 < 0.05
- [6] 0.8647 (0.0 in yes) [8] 0.7728 (0.0 in yes) > 0.01

Step 5 Summarize the results.

- The average rating of booking people and not booking people is equal.

Interpretation:

1. In this case, at the first t-test, I reject null hypothesis because p-value > 0.05
2. I realize that there is only one 0.0 in no sample, so I decide to do it 10 times and see the tendency.
3. Every time there is 0.0 in one side the p-values get higher. However, I do it 10 times and to decide the side which is dominant.

4. There were 3 cases of 10 t-test whose p-value is over 0.01. There were 7 cases of 10 t-test whose p-value is less than 0.01.
5. Therefore, I decide to accept the null hypothesis that 'the average rating of booking people and not booking people is equal.'

T-TEST FOR HYPOTHESIS 4

Step 1 Hypothesis.

The average rating of online order and no-online order is equal.

- H0: μ1 = μ2 (claim) and H1: μ1 ≠ μ2
- H0: μ1 - μ2 = 0 (claim) and H1: μ1 ≠ μ2

N=30 SAMPLING

```
 Descriptive statistics by group 
online: no
        vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
rating     1 30 3.78 1.06      4    4.03 0.3   0 4.5   4.5 -3.02     8.04 0.19
online*    2 30 1.00 0.00      1    1.00 0.0   1 1.0   0.0   NaN      NaN 0.00
-------------------------------------------------------------------------- 
online: yes
        vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
rating     1 30 3.92 0.32      4    3.95 0.3 3.1 4.4   1.3 -0.79     0.57 0.06
online*    2 30 2.00 0.00      2    2.00 0.0 2.0 2.0   0.0   NaN      NaN 0.00
```

Figure 6 descriptive analysis of n=30 sampling of yes or no online-ordered

Step 2 Find the critical value.

- [1] α = 0.05, one-tailed t-value is 4.88. (0.0 in no)

Step 3 Compute the test value.

- [1] CI is 0.25, 0.60 with t-test (alternative="greater").

Step 4 Make the decision.

- [1] Reject the null hypothesis and accept alternative hypothesis. Since 4.88 > 0.6
- [1] 4.88 (0.0 in no) [2] 4.09 (0.0 in no) [3] 4.09 (two 0.0 in no) [4] 4.09 (0.0 in no) [5] 4.09 (two 0.0 in no) > critical t-value

Step 5 Summarize the results.

- The average rating of booking people and not booking people is not equal.

Interpretation:

1. In this case, every sample of online_order 'no' has at least one 0.0 in every t-test.
2. I decided to do it 10 times, but I changed my mind that If I do t-test 5 times with random sampling, every sampling has at least one 0.0 rating in 'No' category. Then, the possibility of reject null hypothesis is very high.
3. My conclusion: reject null hypothesis, because of 5 consecutive same results from different sampling.

T-TEST FOR HYPOTHESIS 5

Step 1 Hypothesis.

The average rating of casual dining and quick bites is same.

- H0: μ1 = μ2 (claim) and H1: μ1 ≠ μ2
- H0: μ1 - μ2 = 0 (claim) and H1: μ1 ≠ μ2

N=30 SAMPLING

```
 Descriptive statistics by group
rest_type: Casual Dining
           vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
rating        1 30 3.68 0.77   3.85    3.81 0.22   0 4.5   4.5 -3.6    14.52 0.14
rest_type*    2 30 1.00 0.00   1.00    1.00 0.00   1 1.0   0.0  NaN      NaN 0.00
------------------------------------------------------------------------
rest_type: Quick Bites
           vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
rating        1 30 3.78 0.33   3.85    3.82 0.22 2.9 4.3   1.4 -1.02     0.35 0.06
rest_type*    2 30 2.00 0.00   2.00    2.00 0.00 2.0 2.0   0.0   NaN      NaN 0.00
```

Figure 7 descriptive analysis of n=30 sampling by two restaurant types

Step 2 Find the critical value.

- α = 0.05, one-tailed p-value is 0.5182

Step 3 Compute the test value.

- p-value is 0.05 with t-test (alternative="two-sided").

Step 4 Make the decision.

- Accept null hypothesis and accept claim. Since 0.5182 > 0.05

Step 5 Summarize the results.

- The average rating of casual dining and quick bites is same.

Interpretation:

1. Until now, I calculated the ratio of two sample variance and see the ratio is less than 4. I use the rule of Thumb which is 'if the ratio of the larger variance to the smaller variance is less than 4 then we can assume the variances are approximately equal (Zach, 2021)'
2. There is no significant difference between Casual Dining restaurant type and Quick Bites restaurant type.

T-TEST FOR HYPOTHESIS 6

Step 1 Hypothesis.

There is a significant difference between Banashankari and Indiranagar in the average of rating.

- H0: μ1 = μ2 and H1: μ1 ≠ μ2 (claim)
- H0: μ1 - μ2 = 0 and H1: μ1 ≠ μ2 (claim)

N=30 SAMPLING

```
 Descriptive statistics by group
location: Banashankari
          vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
rating       1 30  3.9 0.32   3.85    3.91 0.37   3 4.6   1.6 -0.41     0.38 0.06
location*    2 30  1.0 0.00   1.00    1.00 0.00   1 1.0   0.0   NaN      NaN 0.00
-------------------------------------------------------------------------
location: Indiranagar
          vars  n mean   sd median trimmed mad min max range skew kurtosis   se
rating       1 30 4.02 0.39   4.05    4.06 0.3 2.4 4.6   2.2 -2.3     7.82 0.07
location*    2 30 2.00 0.00   2.00    2.00 0.0 2.0 2.0   0.0  NaN      NaN 0.00
```

Figure 8 descriptive analysis of n=30 sampling by two locations

Step 2 Find the critical value.

- α = 0.05, two-tailed p-value is 0.1846

Step 3 Compute the test value.

- p-value is 0.05 with t-test (alternative="two-sided").

Step 4 Make the decision.

- Accept null hypothesis and reject claim. Since 0.1846 > 0.05

Step 5 Summarize the results.

- The average rating which is location by Banashankari and Indiranagar is equal.

Interpretation:

1. There is no significant difference between restaurants located in Banashankari and restaurants located in Indiranagar.

Conclusion

I analyzed the ZOMATO dataset from kaggle. Fortunately, the ZOMATO data contains various categorical data along with the target data (rating), so I was able to do various t-tests. Most interesting part was deciding whether to include a 0.0 rating. I decided to include 0.0, and this decision influenced the two-sample t-test a lot. In the real world, I had to discuss, find out why, and decide how to deal with zero data. I will keep caution about this aspect when analyzing the data in the future.

REFERENCE

Bluman, Allan. (2017). Elementary statistics: a step by step approach 10th edition. McGraw-Hill.

Pranav Uikey & Rushikesh Konapure. (2022). Zomato_EDA. kaggle. Retrieved from https://www.kaggle.com/datasets/pranavuikey/zomato-eda

SANSKRUTI KUNJIR. (2022). EDA_notebook(sanskruti kunjir). kaggle. Retrieved from https://www.kaggle.com/code/sanskrutikunjir/eda-notebook-sanskruti-kunjir

idrrio. (2022, September 29). Describe.by: basic summary statistics by group. Retrieved from https://rdrr.io/cran/psych/man/describe.by.html

STHDA. (n.d.). Unpaired two-samples t-test in R. Retrieved from http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

RDocumentation. (n.d.). T.test: student's t-test. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

ZACH. (2021, April 11). How to Determine equal or uequal variance in t-tests. STATOLOGY. Retrieved from https://www.statology.org/determine-equal-or-unequal-variance/

R-Codes

```
#install packages
install.packages("MASS")
install.packages("psych")
install.packages("stats")
library(easypackages)
libraries("MASS", "psych", "ggpubr", "gplots", "graphics")

#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop")

#Import dataset using read.csv()
zmt <- read.csv("zomato.python cleand.csv", stringsAsFactors = T,
                header=T)

#Checking dataset & data structure
headtail(zmt, 5)
str(zmt)

#rating analysis
table(zmt$rating)
opar <- par(no.readonly = TRUE)
par(fig=c(0, 0.5, 0, 1))
boxplot(zmt$rating, col='#cb202d')
text(y=fivenum(zmt$rating), labels=fivenum(zmt$rating), x=1.35, cex=1)
fivenum(zmt$rating)

#rating histogram
ggplot(zmt, aes(x=rating)) +
  geom_bar()

#subset by cost
high.cost <- subset(zmt, cost > 2000,
           select = c("rating", "cost"))
low.cost <- subset(zmt, cost < 1000,
                select = c("rating", "cost"))
headtail(high.cost,5)
headtail(low.cost,5)

#preparing merge data
df.h <- data.frame(high.cost,"high")
df.h
df.l <- data.frame(low.cost,"low")
df.l
```

```
#change colnames
colnames(df.h)[which(names(df.h) == "X.high.")] <- "cost.range"
colnames(df.l)[which(names(df.l) == "X.low.")] <- "cost.range"

#merge the dataset
df <- rbind(df.h, df.l)
df

# descriptive data
describeBy(df,list(cost.range=df$cost.range))

str(df)
df$cost.range <- as.factor(df$cost.range)

# boxplot of two
plot(rating ~ cost.range, data = df)
mtext("Rating by cost.range", side=3, line=1, cex=2)

#overlay plot
table(df$rating)
plot(x="rating",
    y="Density",
    xlim=range(0:5),
    ylim=range(0:2))
lines(density(df$rating), col = "black")
lines(density(df.h$rating), col = "green")
lines(density(df.l$rating), col = "red", lty=2)
legend(0, 1, legend=c("Total", "low", "high"),
      col=c("black", "red", "green"), lty=c(1,1,2), cex=0.8)

#extract sample 30
sam.df.l <- sample(x=df.l$rating, size=30)
sam.df.l
sam.df.h <- sample(x=df.h$rating, size=30)
sam.df.h

#making data frame
df.sam.l <- data.frame(sam.df.l,"low")
df.sam.l
df.sam.h <- data.frame(sam.df.h,"high")
df.sam.h

#change colnames
colnames(df.sam.l)[which(names(df.sam.l) == "X.low.")] <- "cost.range"
colnames(df.sam.h)[which(names(df.sam.h) == "X.high.")] <- "cost.range"
colnames(df.sam.l)[which(names(df.sam.l) == "sam.df.l")] <- "rating"
```

```
colnames(df.sam.h)[which(names(df.sam.h) == "sam.df.h")] <- "rating"

#rbind
df.sam <- rbind(df.sam.l, df.sam.h)
df.sam

#describeby
df.sam$cost.range <- as.factor(df.sam$cost.range)
describeBy(df.sam,list(cost.range=df.sam$cost.range))


#t-test
t.test(sam.df.h, sam.df.l, alternative = "greater", var.equal = TRUE)

#subset by votes
high.votes <- subset(zmt, votes > 5000,
                  select = c("rating", "votes"))
low.votes <- subset(zmt, cost < 2500,
                 select = c("rating", "votes"))
headtail(high.votes,5)
headtail(low.votes,5)

#extract sample 30
v.df.h <- sample(x=high.votes$rating, size=30)
v.df.h
v.df.l <- sample(x=low.votes$rating, size=30)
v.df.l

#making data frame
v.sam.l <- data.frame(v.df.l,"less")
v.sam.l
v.sam.h <- data.frame(v.df.h,"many")
v.sam.h

#change colnames
colnames(v.sam.l)[which(names(v.sam.l) == "X.less.")] <- "votes"
colnames(v.sam.h)[which(names(v.sam.h) == "X.many.")] <- "votes"
colnames(v.sam.l)[which(names(v.sam.l) == "v.df.l")] <- "rating"
colnames(v.sam.h)[which(names(v.sam.h) == "v.df.h")] <- "rating"

#rbind
df.v <- rbind(v.sam.l, v.sam.h)
df.v

#describeby
df.v$votes <- as.factor(df.v$votes)
```

```
describeBy(df.v,list(votes=df.v$votes))


#t-test
t.test(v.df.h, v.df.l, alternative = "greater", var.equal = TRUE)

str(zmt)
## HYPOTHESIS 3
#subset by votes
book.yes <- subset(zmt, booking == "Yes",
                   select = c("rating", "booking"))
book.no <- subset(zmt, booking == "No",
                  select = c("rating", "booking"))
headtail(book.yes,5)
headtail(book.no,5)

#extract sample 30
b.df.y <- sample(x=book.yes$rating, size=30)
b.df.y
b.df.n <- sample(x=book.no$rating, size=30)
b.df.n

#making data frame
b.sam.y <- data.frame(b.df.y,"yes")
b.sam.y
b.sam.n <- data.frame(b.df.n,"no")
b.sam.n

#change colnames
colnames(b.sam.y)[which(names(b.sam.y) == "X.yes.")] <- "booking"
colnames(b.sam.n)[which(names(b.sam.n) == "X.no.")] <- "booking"
colnames(b.sam.y)[which(names(b.sam.y) == "b.df.y")] <- "rating"
colnames(b.sam.n)[which(names(b.sam.n) == "b.df.n")] <- "rating"
b.sam.y
b.sam.n

#rbind
df.b <- rbind(b.sam.y, b.sam.n)
df.b

#describeby
df.b$booking <- as.factor(df.b$booking)
describeBy(df.b,list(booking=df.b$booking))


#t-test
```

```
t.test(b.df.y, b.df.n, alternative = "two.sided", var.equal = TRUE)


str(zmt)
## HYPOTHESIS 4
#subset by online_order
online.yes <- subset(zmt, online_order == "Yes",
                  select = c("rating", "booking"))
online.no <- subset(zmt, online_order == "No",
                 select = c("rating", "booking"))
headtail(online.yes,5)
headtail(online.no,5)

#extract sample 30
o.df.y <- sample(x=online.yes$rating, size=30)
o.df.y
o.df.n <- sample(x=online.no$rating, size=30)
o.df.n

#making data frame
o.sam.y <- data.frame(o.df.y,"yes")
o.sam.y
o.sam.n <- data.frame(o.df.n,"no")
o.sam.n

#change colnames
colnames(o.sam.y)[which(names(o.sam.y) == "X.yes.")] <- "online"
colnames(o.sam.n)[which(names(o.sam.n) == "X.no.")] <- "online"
colnames(o.sam.y)[which(names(o.sam.y) == "o.df.y")] <- "rating"
colnames(o.sam.n)[which(names(o.sam.n) == "o.df.n")] <- "rating"
o.sam.y
o.sam.n

#rbind
df.o <- rbind(o.sam.y, o.sam.n)
df.o

#describeby
df.o$online <- as.factor(df.o$online)
describeBy(df.o,list(online=df.o$online))


#t-test
t.test(b.df.y, b.df.n, alternative = "two.sided", var.equal = TRUE)

str(zmt)
```

```
table(zmt$rest_type)

## HYPOTHESIS 5
#subset by restaurant type
rest_type.casu <- subset(zmt, rest_type == "Casual Dining",
                    select = c("rating", "rest_type"))
rest_type.quick <- subset(zmt, rest_type == "Quick Bites",
                   select = c("rating", "rest_type"))
headtail(rest_type.casu,5)
headtail(rest_type.quick,5)

#extract sample 30
t.df.c <- sample(x=rest_type.casu$rating, size=30)
t.df.c
t.df.q <- sample(x=rest_type.quick$rating, size=30)
t.df.q

#making data frame
t.sam.c <- data.frame(t.df.c,"Casual Dining")
t.sam.c
t.sam.q <- data.frame(t.df.q,"Quick Bites")
t.sam.q

#change colnames
colnames(t.sam.c)[which(names(t.sam.c) == "X.Casual.Dining.")] <- "rest_type"
colnames(t.sam.q)[which(names(t.sam.q) == "X.Quick.Bites.")] <- "rest_type"
colnames(t.sam.c)[which(names(t.sam.c) == "t.df.c")] <- "rating"
colnames(t.sam.q)[which(names(t.sam.q) == "t.df.q")] <- "rating"
t.sam.c
t.sam.q

#rbind
df.t <- rbind(t.sam.c, t.sam.q)
df.t

#describeby
df.t$rest_type <- as.factor(df.t$rest_type)
describeBy(df.t,list(rest_type=df.t$rest_type))


#t-test
t.test(t.df.c, t.df.q, alternative = "two.sided", var.equal = TRUE)


str(zmt)
table(zmt$location)
```

```
## HYPOTHESIS 6
#subset by location
loca.ba <- subset(zmt, location == "Banashankari",
                   select = c("rating", "location"))
loca.in <- subset(zmt, location == "Indiranagar",
                   select = c("rating", "location"))
headtail(loca.ba,5)
headtail(loca.in,5)

#extract sample 30
l.df.b <- sample(x=loca.ba$rating, size=30)
l.df.b
l.df.i <- sample(x=loca.in$rating, size=30)
l.df.i

#making data frame
l.sam.b <- data.frame(l.df.b,"Banashankari")
l.sam.b
l.sam.i <- data.frame(l.df.i,"Indiranagar")
l.sam.i

#change colnames
colnames(l.sam.b)[which(names(l.sam.b) == "X.Banashankari.")] <- "location"
colnames(l.sam.i)[which(names(l.sam.i) == "X.Indiranagar.")] <- "location"
colnames(l.sam.b)[which(names(l.sam.b) == "l.df.b")] <- "rating"
colnames(l.sam.i)[which(names(l.sam.i) == "l.df.i")] <- "rating"
l.sam.b
l.sam.i

#rbind
df.l <- rbind(l.sam.b, l.sam.i)
df.l

#describeby
df.l$location <- as.factor(df.l$location)
describeBy(df.l,list(location=df.l$location))


#t-test
t.test(l.df.b, l.df.i, alternative = "two.sided", var.equal = TRUE)
```