Exercise-4

Jamia Begum

2023-02-13

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   0.3.5
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

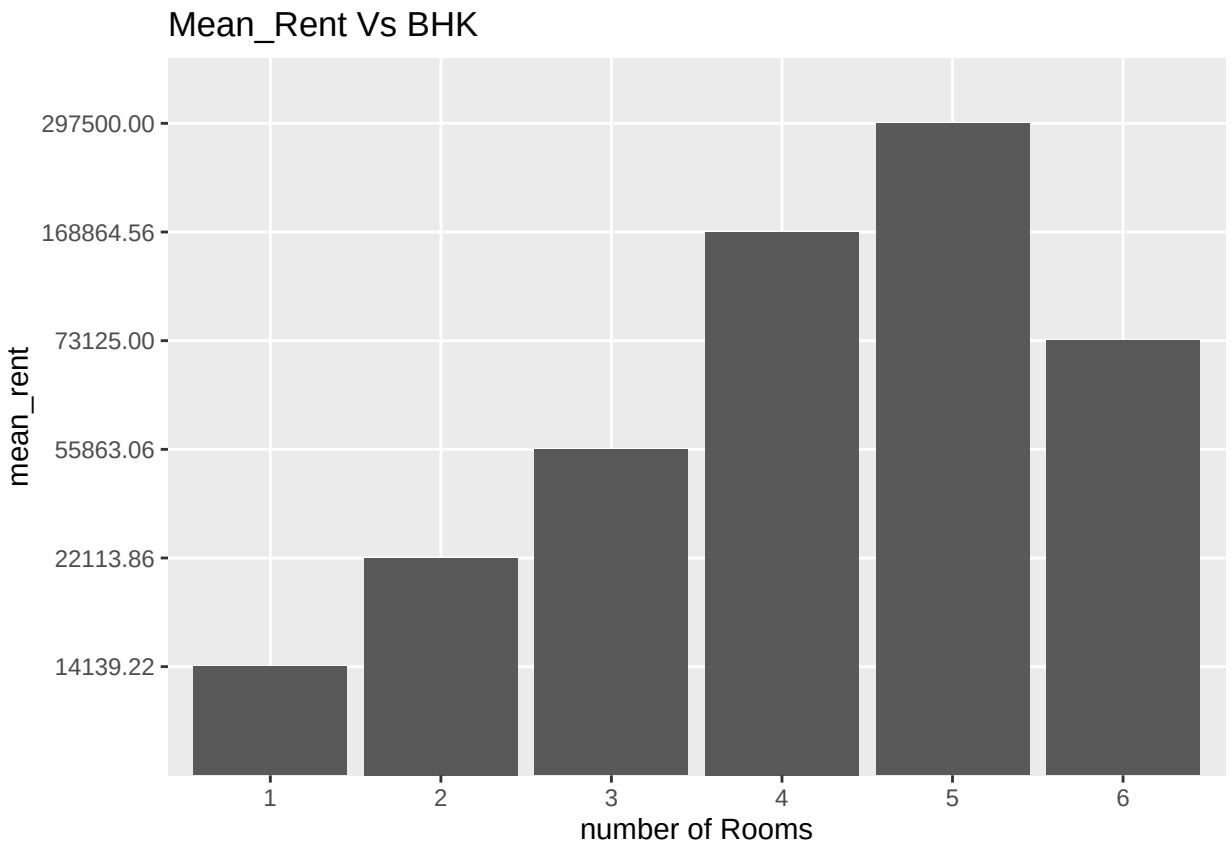
```
house<- read.csv("F:/INTERMATH/intermath 2021-2023/spain/DV/Resampling/Ex-4/rent bootsrap for india/arc
view(house) #to see the dataset
```

#1.How the Rent of the houses varies according to various variables such as: the Number of Bedrooms and Hall/Kitchen areas (BHK),City, and Furnishing status.

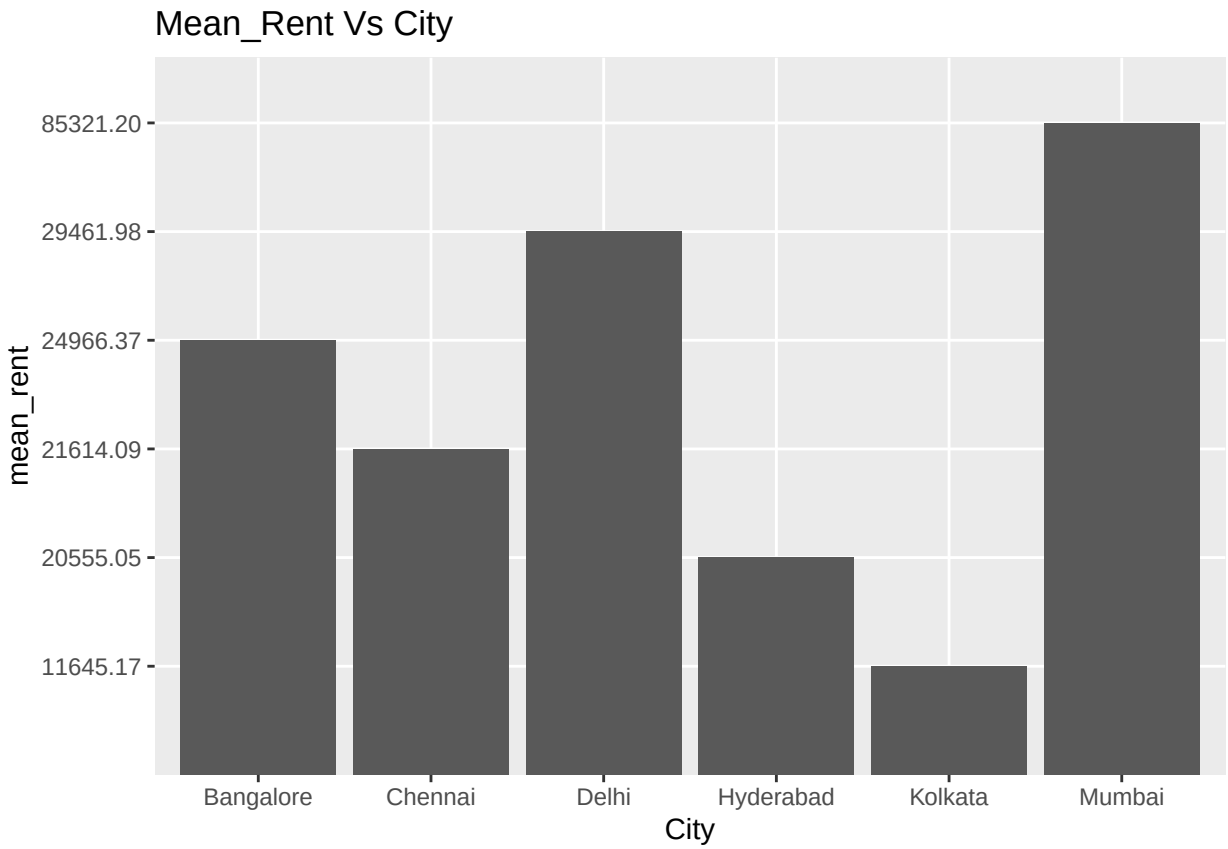
```
# mean_rent is defined as the mean rent of houses grouped by BHK categories
meanrent_by_rooms<- house %>%
  group_by(BHK) %>%
  summarize(mean_rent = mean(Rent))
meanrent_by_rooms
```

```
## # A tibble: 6 x 2
##     BHK mean_rent
##   <int>     <dbl>
## 1     1    14139.
## 2     2    22114.
## 3     3    55863.
## 4     4   168865.
## 5     5   297500
## 6     6    73125
```

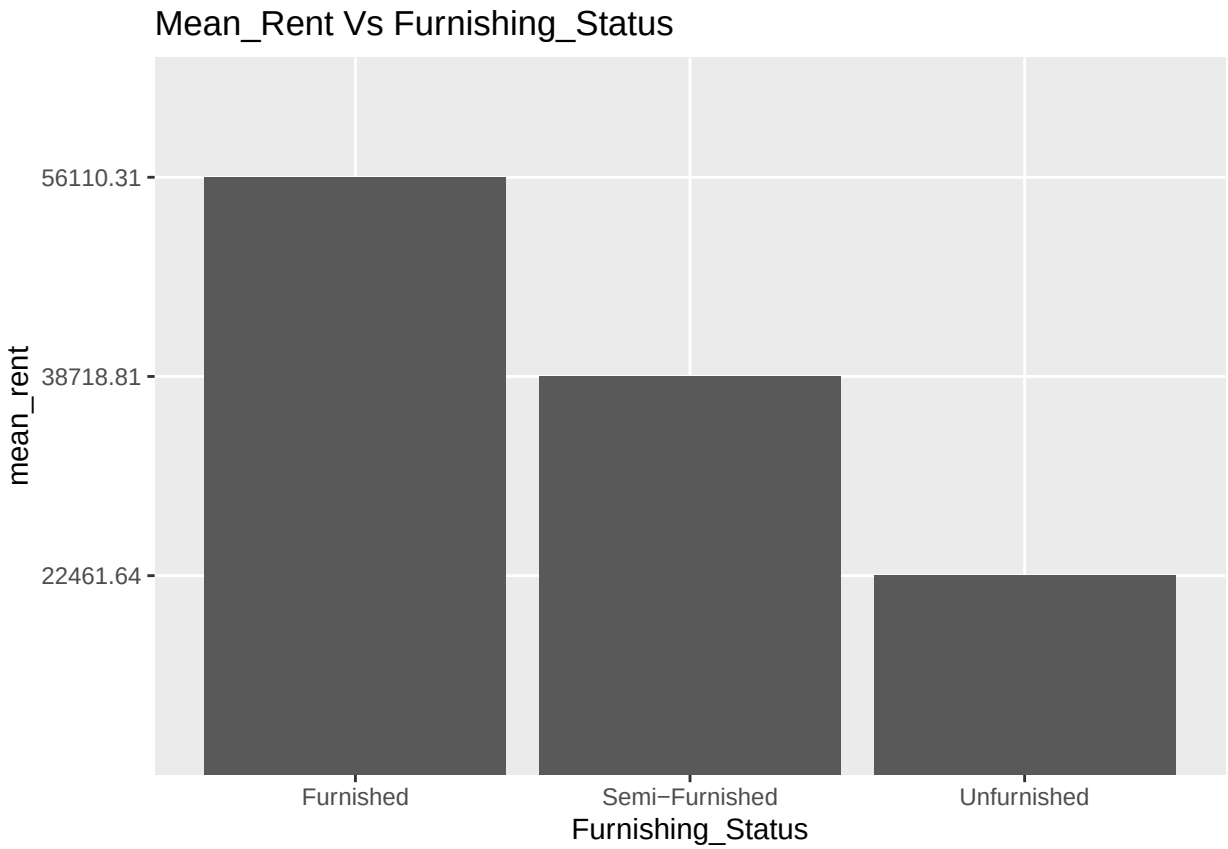
```
meanrent_by_rooms%>%
ggplot(aes(x = factor(BHK), y =format(mean_rent ,scientific=FALSE))) +
  labs(title= "Mean_Rent Vs BHK",x="number of Rooms",y="mean_rent")+
 geom_bar(stat = "identity")
```




```
house %>%
 group_by(City) %>%
 summarize(mean_rent = mean(Rent))%>%
ggplot(aes(x = City, y =format(mean_rent ,scientific=FALSE))) +
  labs(title= "Mean_Rent Vs City",x="City",y="mean_rent")+
 geom_bar(stat = "identity")
```

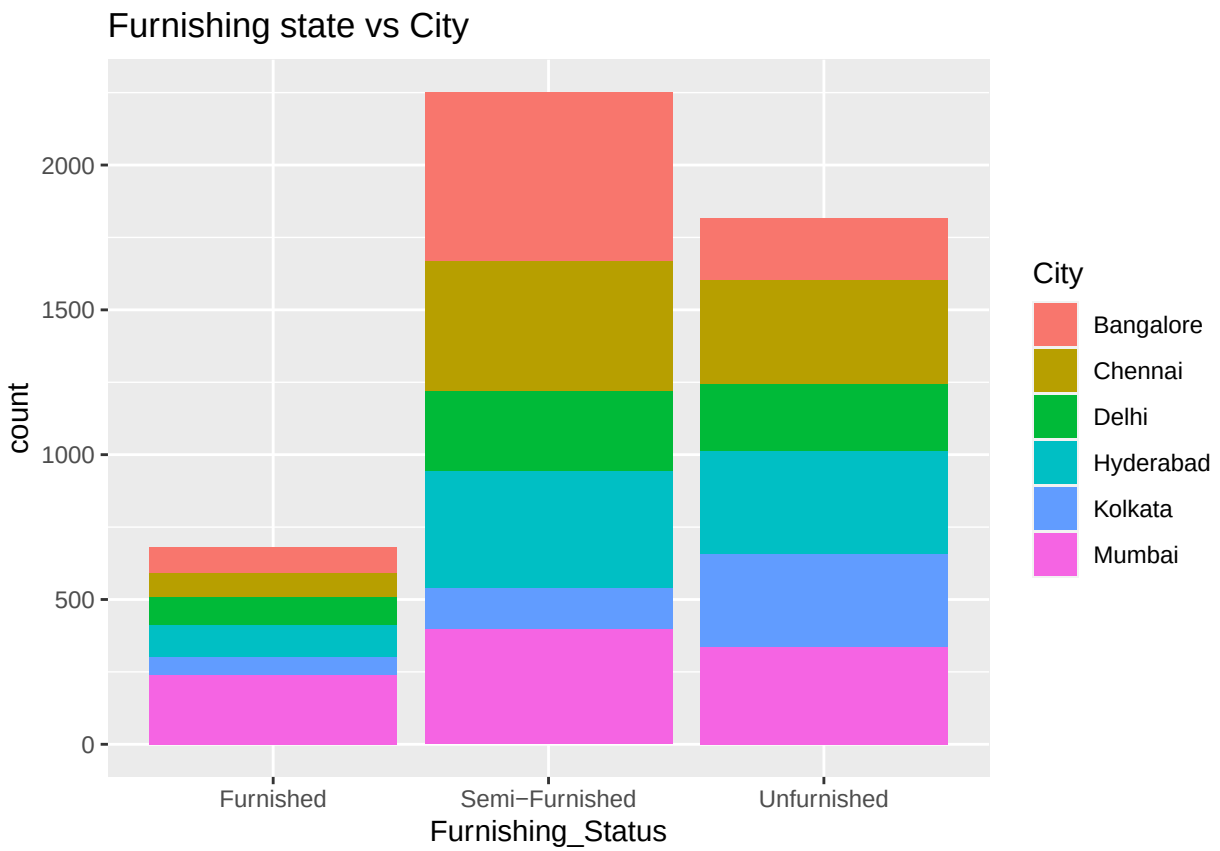



```
house %>%
 group_by(Furnishing_Status) %>%
 summarize(mean_rent = mean(Rent))%>%
ggplot(aes(x = Furnishing_Status, y =format(mean_rent ,scientific=FALSE))) +
  labs(title= "Mean_Rent Vs Furnishing_Status",x="Furnishing_Status",y="mean_rent")+
 geom_bar(stat = "identity")
```



#2.How is the furnishing state of the houses according to the cities?

```
house %>%
  ggplot() +
  geom_bar(mapping = aes(x = Furnishing_Status,fill=City))+
  labs(title="Furnishing state vs City")
```

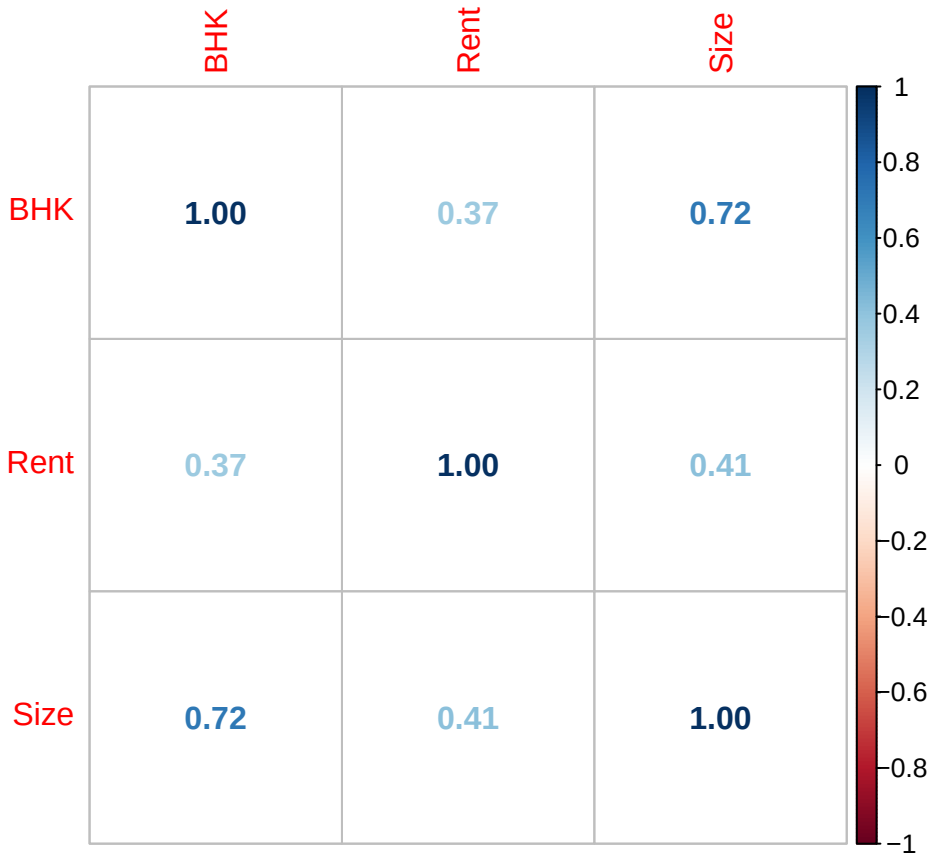


#3. Show the correlation between the numerical variables BHK, Rent, and Size

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

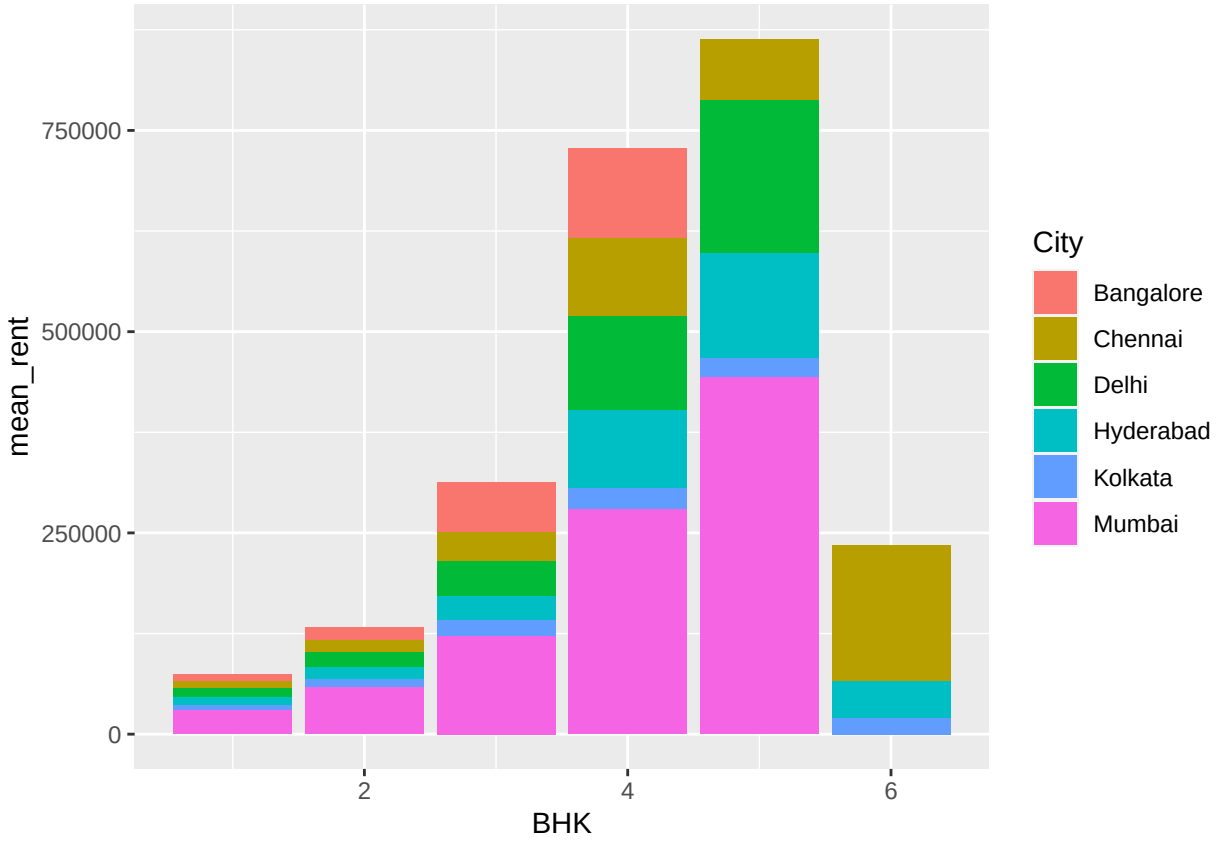
```
M <- cor(house[,c(2,3,4)])
corrplot(M, method = 'number')
```



#4.How rent price varies with respect to the City and BHK?

```
house %>%
  group_by(BHK,City) %>%
  summarize(mean_rent = mean(Rent)) %>%
  ggplot(aes(x = BHK, y = mean_rent, group = City, fill = City)) +
  geom_bar(stat = "identity")
```

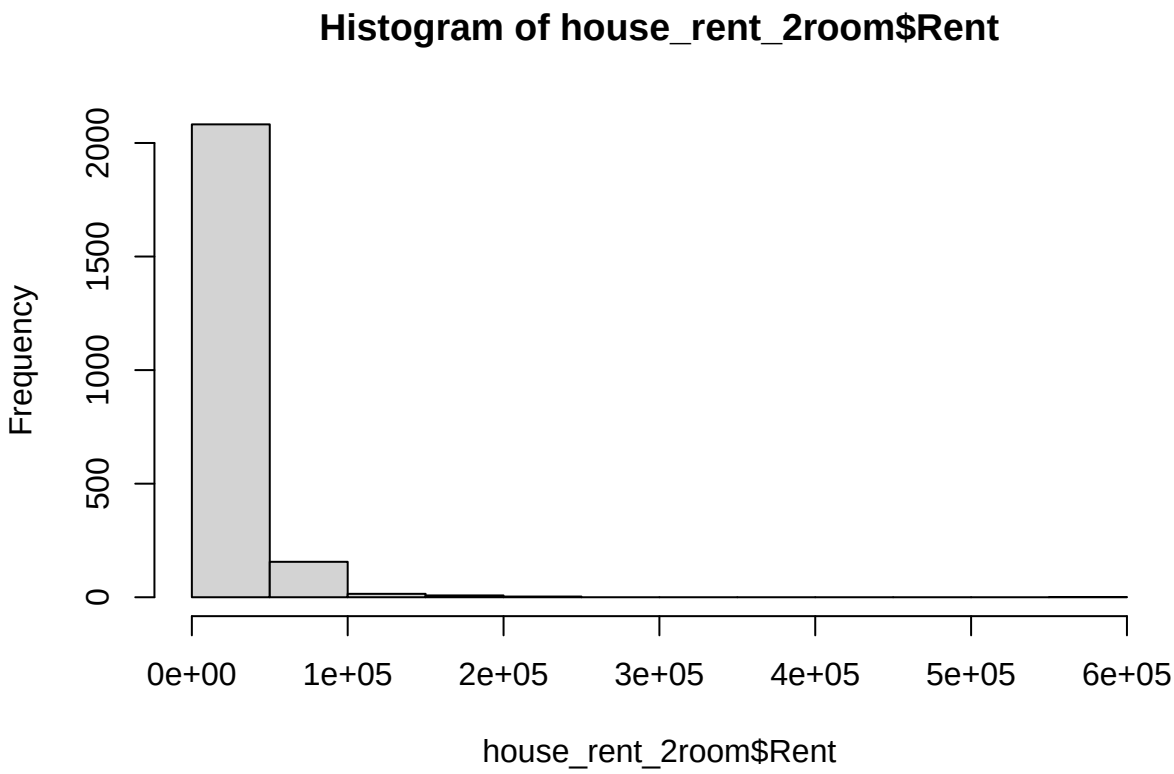
```
## `summarise()` has grouped output by 'BHK'. You can override using the `.groups`
## argument.
```



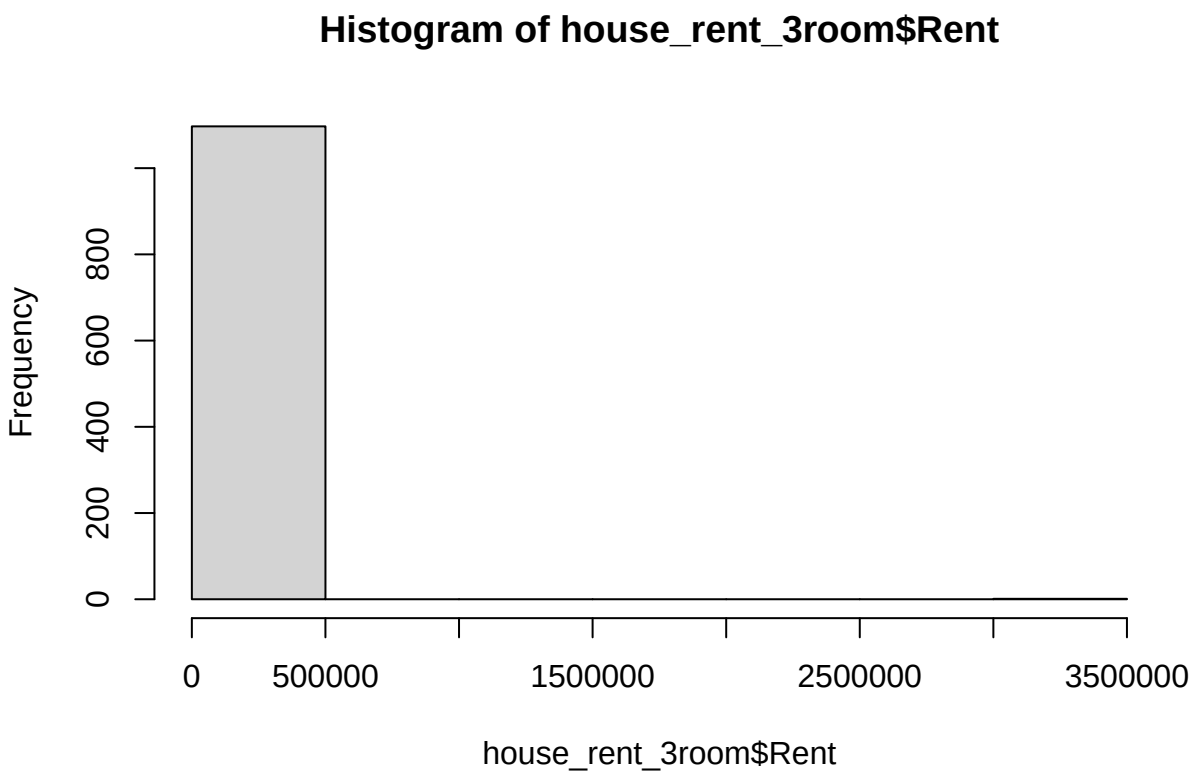
#5. i)What is a better measure for typical rent in India for both 2-room and 3-room houses?

```
#select rent data according to the no. of rooms
house_rent<- house %>% select("BHK","Rent")
#select rent data for 2 rooms house
house_rent_2room <- house_rent[house_rent$BHK=="2",]
#select rent data for 3 rooms house
house_rent_3room<- house_rent[house_rent$BHK=="3",]
```

```
hist(house_rent_2room$Rent)
```




```
hist(house_rent_3room$Rent)
```




```
#Since in both of our cases, we don't have any outliers,
#we will use mean as the measure for typical rents
```

#5.ii) Use bootstrap techniques to estimate the mean rental price for the whole population in India

```
library(tidyverse)
#bootstrapping using infer package
library(infer)
```

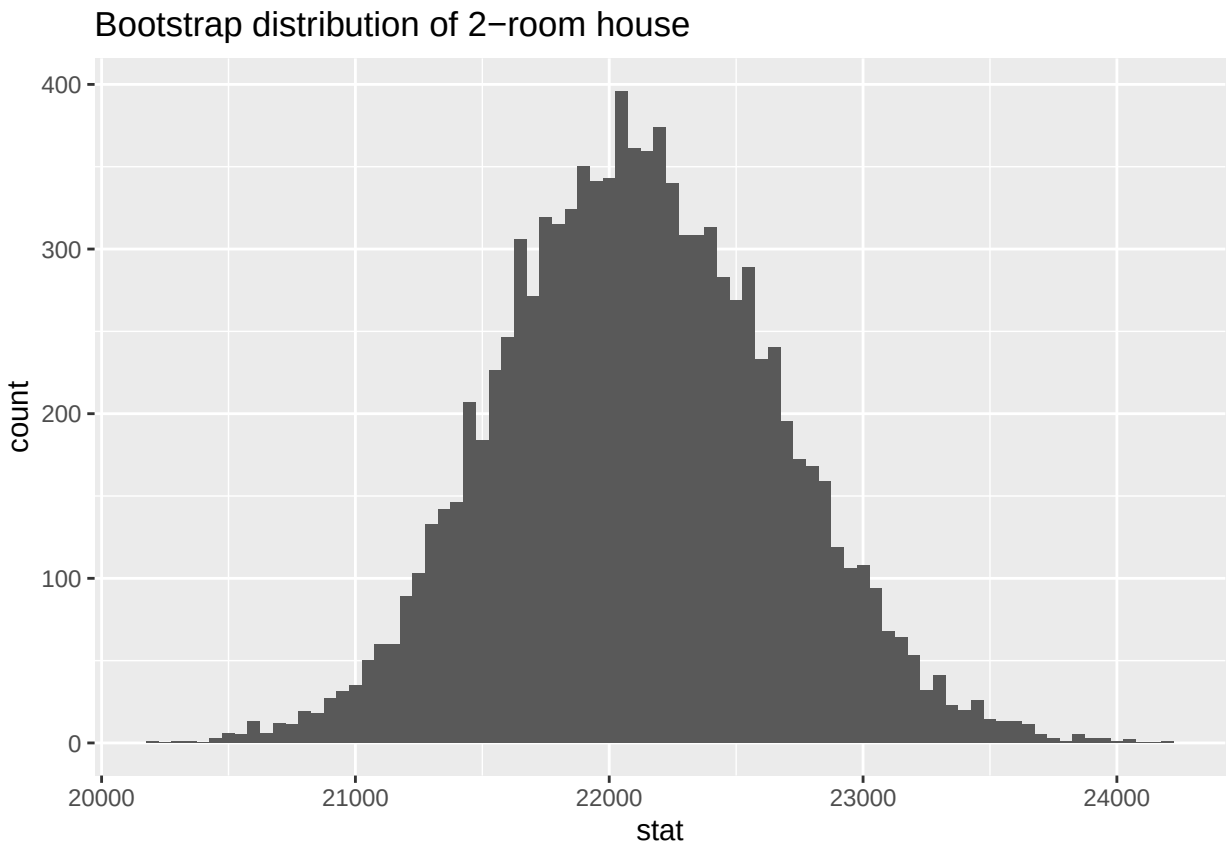
```
## Warning: package 'infer' was built under R version 4.2.2
```

```
# Generate bootstrap distribution of means for 2-room houses:
set.seed(1)
rent_mean2 <- house_rent_2room %>%
  # Specify the variable of interest
specify(response = Rent) %>%
  # Generate 10000 bootstrap samples
  generate(reps = 10000, type = "bootstrap") %>%
  # Calculate the mean of each bootstrap sample
  calculate(stat = "mean")

# View its structure
str(rent_mean2)
```

```
## infer [10,000 x 2] (S3: infer/tbl_df/tbl/data.frame)
```

```
##  $ replicate: int [1:10000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ stat     : num [1:10000] 22870 22106 21931 23761 22394 ...
##  - attr(*, "response")= symbol Rent
##  - attr(*, "response_type")= chr "numeric"
##  - attr(*, "distr_param")= Named num 2264
##   ..- attr(*, "names")= chr "df"
##  - attr(*, "theory_type")= chr "One sample t"
##  - attr(*, "generated")= logi TRUE
##  - attr(*, "type")= chr "bootstrap"
##  - attr(*, "hypothesized")= logi FALSE
##  - attr(*, "fitted")= logi FALSE
##  - attr(*, "type_desc_response")= chr "num"
##  - attr(*, "type_desc_explanatory")= chr ""
##  - attr(*, "stat")= chr "mean"
```

```
# Plot the rent_mean2 statistic
ggplot(rent_mean2, aes(x=stat)) +
  # Make it a histogram with a binwidth of 50
  geom_histogram(binwidth=50) +
  labs(title = "Bootstrap distribution of 2-room house ")
```

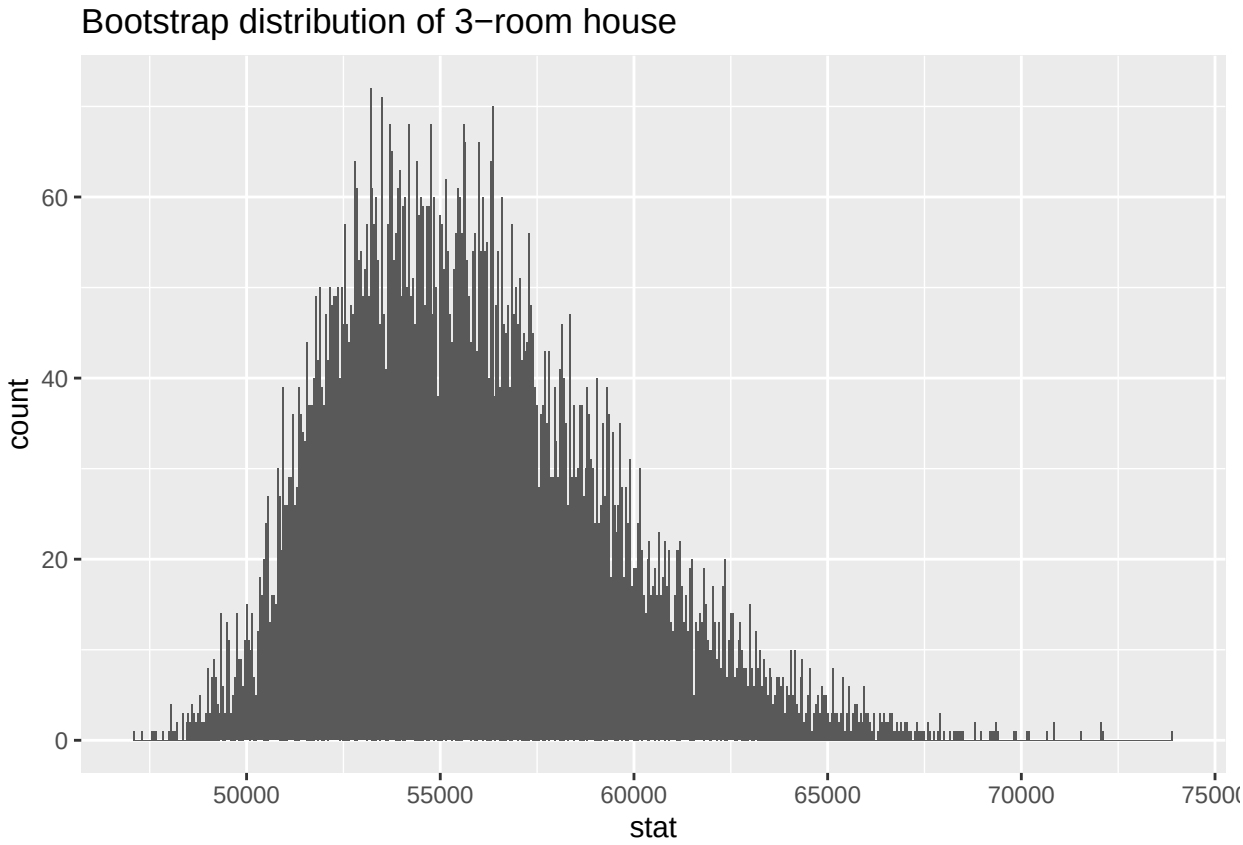



```
# Similarly,generate bootstrap distribution of means for 3-room houses:
set.seed(1)
rent_mean3 <- house_rent_3room %>%
  # Specify the variable of interest
```

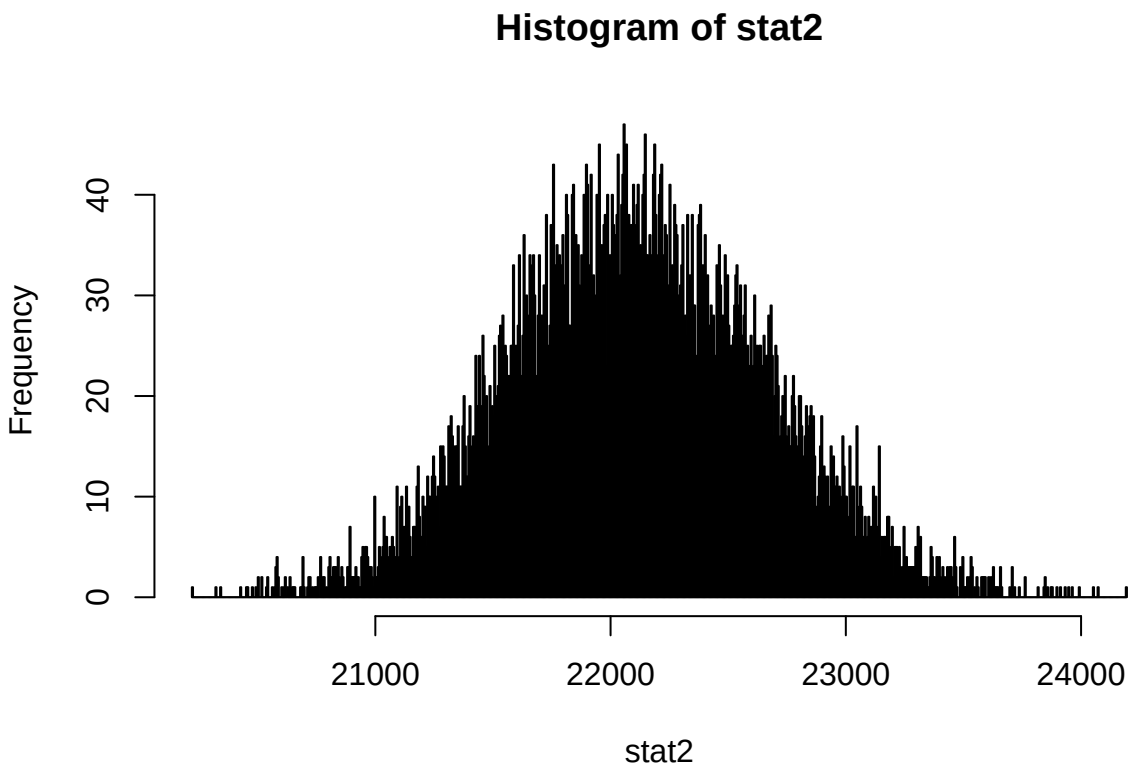
```r
specify(response = Rent) %>%
  # Generate 10000 bootstrap samples
  generate(reps = 10000, type = "bootstrap") %>%
  # Calculate the mean of each bootstrap sample
  calculate(stat = "mean")

# View its structure
str(rent_mean3)
```

```
## infer [10,000 x 2] (S3: infer/tbl_df/tbl/data.frame)
##  $ replicate: int [1:10000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ stat     : num [1:10000] 62715 56896 53407 57176 55994 ...
##  - attr(*, "response")= symbol Rent
##  - attr(*, "response_type")= chr "numeric"
##  - attr(*, "distr_param")= Named num 1097
##   ..- attr(*, "names")= chr "df"
##  - attr(*, "theory_type")= chr "One sample t"
##  - attr(*, "generated")= logi TRUE
##  - attr(*, "type")= chr "bootstrap"
##  - attr(*, "hypothesized")= logi FALSE
##  - attr(*, "fitted")= logi FALSE
##  - attr(*, "type_desc_response")= chr "num"
##  - attr(*, "type_desc_explanatory")= chr ""
##  - attr(*, "stat")= chr "mean"
```

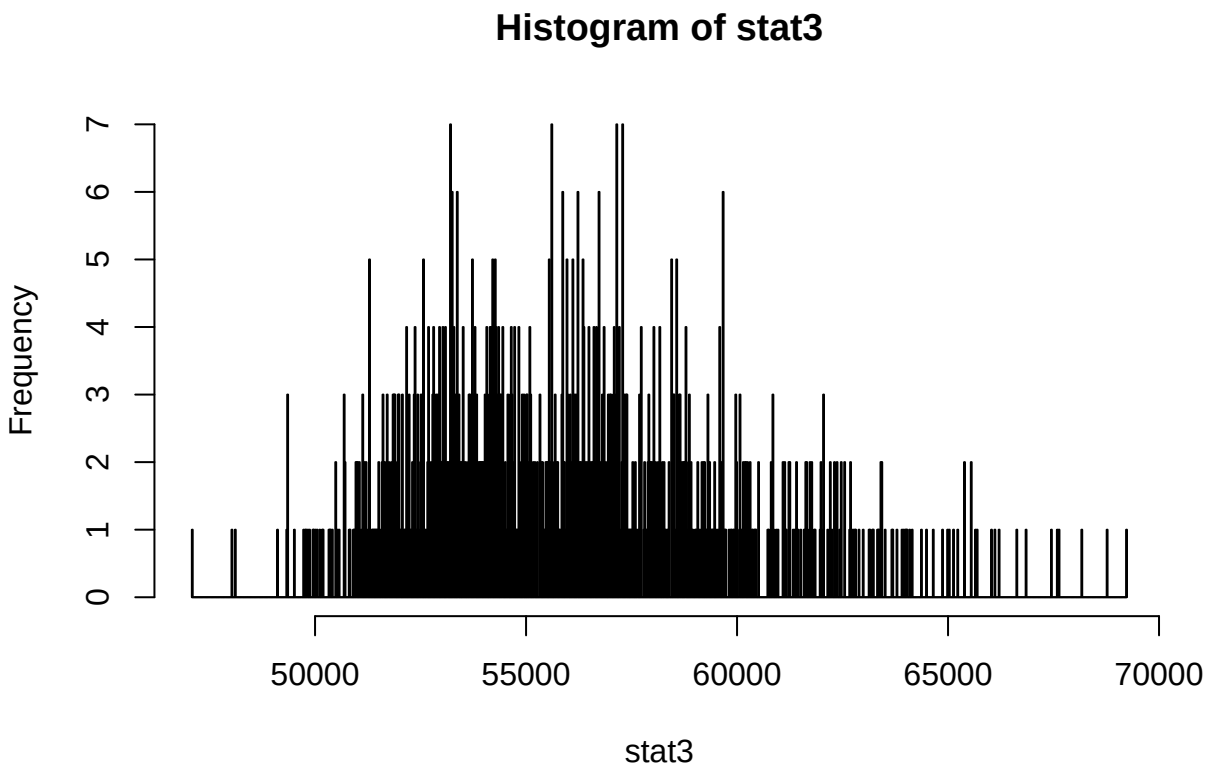
```r
# Plot the rent_mean2 statistic
ggplot(rent_mean3, aes(x=stat)) +
  # Make it a histogram with a binwidth of 50
  geom_histogram(binwidth=50) +
  labs(title = "Bootstrap distribution of 3-room house ")
```




```
#manually bootstraping for 2-room houses:
set.seed(1)
rent2<-house_rent_2room$Rent
n =length(rent2)
#set number of bootstrap samples
nsim =10000
stat2 = numeric(nsim ) #create a vector in which to store the results
#Set up a loop to generate a series of bootstrap samples
for (i in 1:nsim){
rent2B = sample(rent2 , n, replace = T)
stat2[i] = mean(rent2B)}
hist(stat2,breaks=1000)
```




```
#manually bootstraping for 3-room houses:
set.seed(1)
rent3<-house_rent_3room$Rent
n =length(rent3)
#set number of bootstrap samples
nsim =1000
stat3 = numeric(nsim ) #create a vector in which to store the results
se3=numeric(nsim )
#Set up a loop to generate a series of bootstrap samples
for (i in 1:nsim){
rent3B = sample(rent3 , n, replace = T)
stat3[i] = mean(rent3B)
}
hist(stat3,breaks=1000)
```



#percentile method to calculate 95% CI of mean rent for 2-room and 3-room houses

```
CI2<-quantile(stat2,c(0.025,0.975))
#for 2-room house rent
per_ci2<-tibble(lower=CI2[1],upper=CI2[2])
per_ci2
```

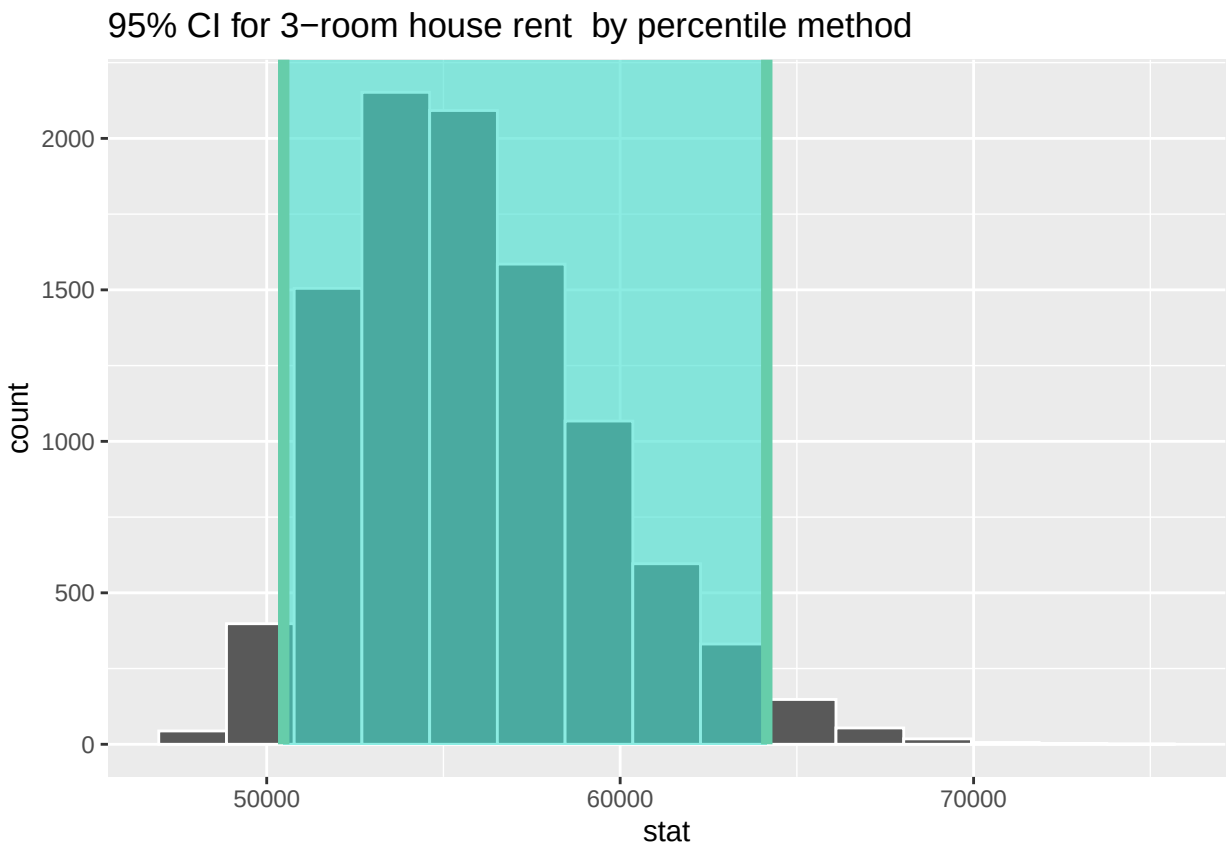
```
## # A tibble: 1 x 2
##    lower  upper
##    <dbl>  <dbl>
## 1 21091. 23189.
```

```
visualize(rent_mean2) +
shade_confidence_interval(endpoints = per_ci2)+
  labs(title = "95% CI for 2-room house rent  by percentile method ")
```


```
CI3<-quantile(stat3,c(0.025,0.975))
#for 3-room house rent
per_ci3<-tibble(lower=CI3[1],upper=CI3[2])
per_ci3
```

```
## # A tibble: 1 x 2
##    lower  upper
##    <dbl>  <dbl>
## 1 50480. 64149.
```

```
visualize(rent_mean3) +
shade_confidence_interval(endpoints = per_ci3)+
  labs(title = "95% CI for 3-room house rent  by percentile method ")
```



#bootstrap-t method to calculate 95% CI of mean rent for 2-room and 3-room houses

```
#estimating 95% CI for 2-room houses using bootstrap-t

rent2<-house_rent_2room$Rent

#rent mean from original sample is the sample statistic here
sample_statistic2<-mean(rent2)
sample_statistic2
```

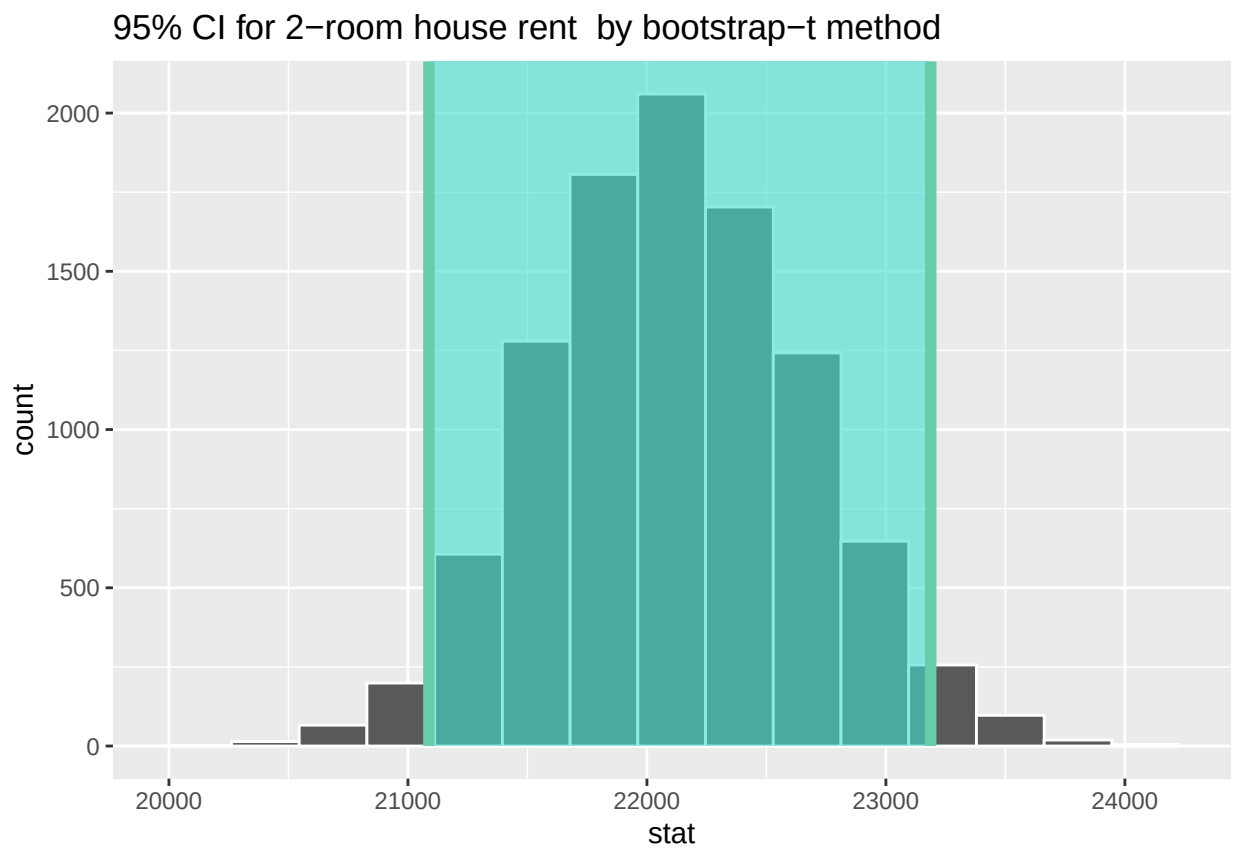
```
## [1] 22113.86
```

```
#bootstrap mean for 2-room house rent is stored in rent_mean2[2]

#calculating bootstrap standard error of the statistic
seb2<-rent_mean2 %>% specify(response = stat) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "sd")


#calculating bootstrap t-values
t2<-(rent_mean2[2]- sample_statistic2)/seb2[2]

# Calculate the std error of the statistic
se2<-rent_mean2 %>%
```

```
  summarize(se = sd(stat)) %>%pull()

# Calculate the lower and upper limits of the CI
    l2 = sample_statistic2 + se2 * quantile(t2[[1]],0.025)
    u2 = sample_statistic2 +se2 * quantile(t2[[1]],0.975)
  tci2<-tibble(lower=l2,upper=u2)
tci2
```

```
## # A tibble: 1 x 2
##    lower  upper
##    <dbl>  <dbl>
## 1 21088. 23188.
```

```
visualize(rent_mean2) +
shade_confidence_interval(endpoints = tci2)+
  labs(title = "95% CI for 2-room house rent  by bootstrap-t method ")
```



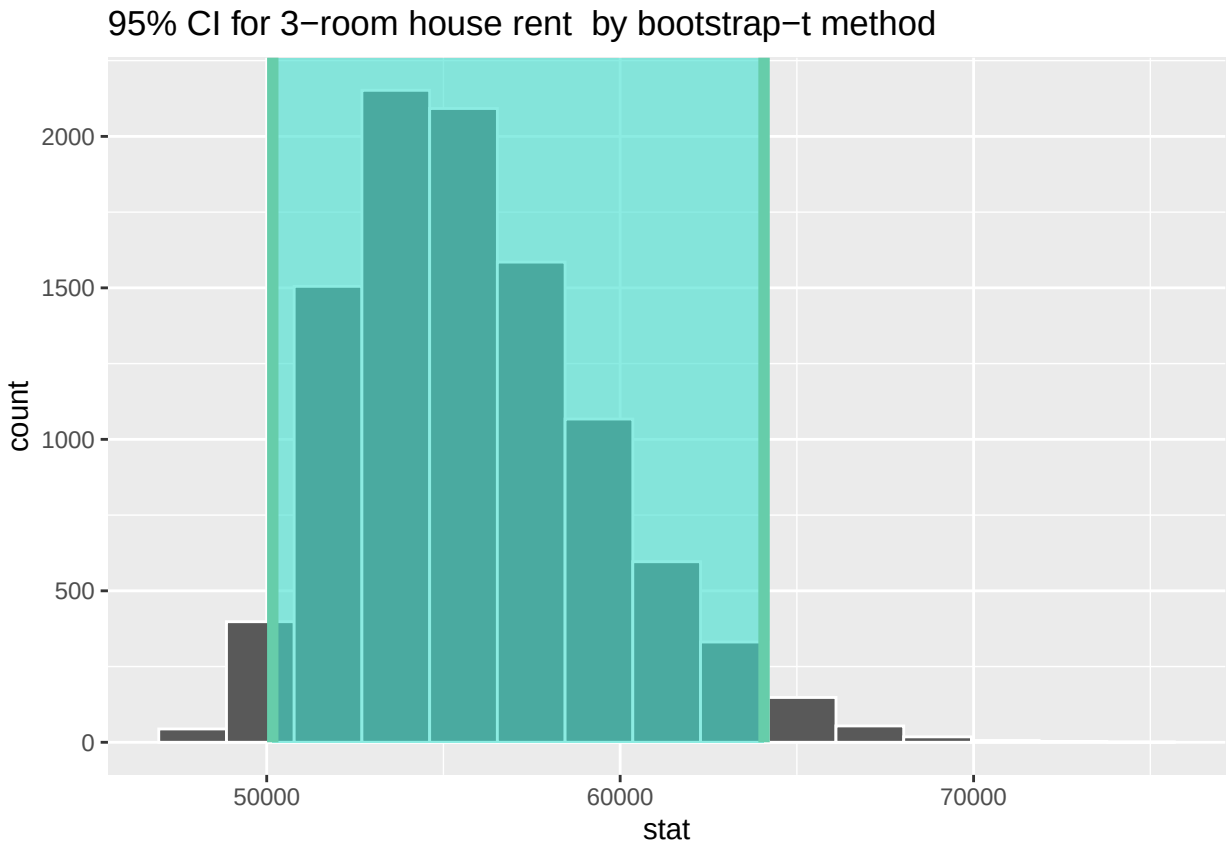

```
#estimating 95% CI for 3-room houses using bootstrap-t

rent3<-house_rent_3room$Rent

#rent mean from observation is the sample statistic here
sample_statistic3<-mean(rent3)
sample_statistic3
```

```
## [1] 55863.06
```

```
#bootstrap mean for 3-room houses is stored in rent_mean3[2]

#calculating bootstrap standard error of the statistic
seb3<-rent_mean3 %>% specify(response = stat) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "sd")

#calculating bootstrap t-values
t3<-(rent_mean3[2]- sample_statistic3)/seb3[2]

# Calculate the std error of the statistic
se3<-rent_mean3 %>%
  summarize(se = sd(stat)) %>%pull()

# Calculate the lower and upper limits of the 95% CI
    l3 = sample_statistic3 + se3 * quantile(t3[[1]],0.025)
    u3 = sample_statistic3 +se3 * quantile(t3[[1]],0.975)
 tci3<-tibble(lower=l3,upper=u3)
tci3
```

```
## # A tibble: 1 x 2
##    lower  upper
##    <dbl>  <dbl>
## 1 50171. 64072.
```

```
visualize(rent_mean3) +
shade_confidence_interval(endpoints = tci3)+
  labs(title = "95% CI for 3-room house rent  by bootstrap-t method ")
```



#5.iii) Evaluate whether this data provides evidence that the mean rent of 2-room houses in India is different than 21000 Rupee?

```
#Calculate 1500 bootstrap replicates of the mean rent.
#Use a point null hypothesis of mean rent being mu = 24000 Rupee.
n_sample <- 1500

rent_mean_ht <- house_rent_2room %>%
  # Specify rent as the response(the variable we want to consider)
  specify(response= Rent) %>%
# Set the point hypothesis that mean is 24000 Rupee(create the null hypothesis)
  hypothesize(null="point",mu= 21000) %>%
  # Generate 1500 bootstrap samples
  generate(reps=n_sample,type="bootstrap") %>%
  # Calculate the mean for each sample
  calculate(stat="mean")


#Calculate the mean rent from the original observed sample
#and pull out the value.
rent_mean_obs <- house_rent_2room %>%
  summarize(mean_rent = mean(Rent)) %>%
  pull()
#Calculate the two-sided p-value.
 rent_mean_ht%>%
  get_p_value(rent_mean_obs, direction = "two-sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0507
```

```
#visualize the null distribution and comparing it to the observed statistic
rent_mean_ht %>%
  visualize() +
  shade_p_value(rent_mean_obs, direction = "two-sided")
```