Assignment-4

Title: Statistical Analysis using Data Visualization Techniques with Resampling

Course: Data Visualization and Modelling

Professor: Pere Puig

Submitted By: Jamia Begum

NIU: 1676891

Abstract: This report presents an analysis of a dataset containing information on 4747 rental properties in India, including houses, apartments, and flats. The dataset has 12 variables, of which 5 are numerical and the remaining are categorical. We used the tydiverse package for data cleaning, transforming, and visualizing data. The ggplot package in the tydiverse package has been used to create informative and visually appealing visualizations. Then we analyzed how rent prices vary according to various variables, such as the number of bedrooms, hall, and kitchens (BHK), cities, and furnishing status. The results showed that the mean rent price is highest for 5-room houses and lowest for 1-room houses, highest in Mumbai and lowest in Kolkata, and highest for furnished houses and lowest for unfurnished houses. Rent prices are highest for all BHK types of houses in Mumbai. Only Chennai, Kolkata, and Hyderabad have 6-BHK houses with comparatively low prices. We also used the corrplot library to show the correlation between the numerical variables of BHK, rent, and size. The size and number of rooms (BHK) of the houses are strongly correlated.

Additionally, we extracted the monthly rent data for 2-room and 3-room apartments in India to estimate the typical rent for both 2-room and 3-room houses in India. We used bootstrap techniques to estimate the mean rental price for the whole population in India. Bootstrapping is a resampling technique that creates multiple datasets from a single sample, and is useful when the sample size is small compared to the population. The non-parametric bootstrap method has been used here, which involves taking multiple random samples with replacement, calculating the sample statistic for each resample, and getting an estimate of the bootstrap statistic from the bootstrap distribution. Infer package and manual bootstrapping both can be used. Two methods were used to calculate confidence intervals: the percentile method and the bootstrap-t method. Both methods provided similar results for both 2-room and 3-room houses. A hypothesis test has been also conducted to determine whether the mean rent of 2-room houses in India is different than 21000 Rupee. The result suggests that there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

Dataset:

The dataset has been taken from kaggle: https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset?select=House_Rent_Dataset.csv

This dataset includes information of 4747 rental properties in India, such as houses, apartments, and flats, with various specifications such as the number of bedrooms, Hall, Kitchen (BHK), Rent, Size, No. of Floors, Area Type, Area Locality, City, Furnishing Status, Type of Tenant Preferred, No. of Bathrooms, Point of Contact.

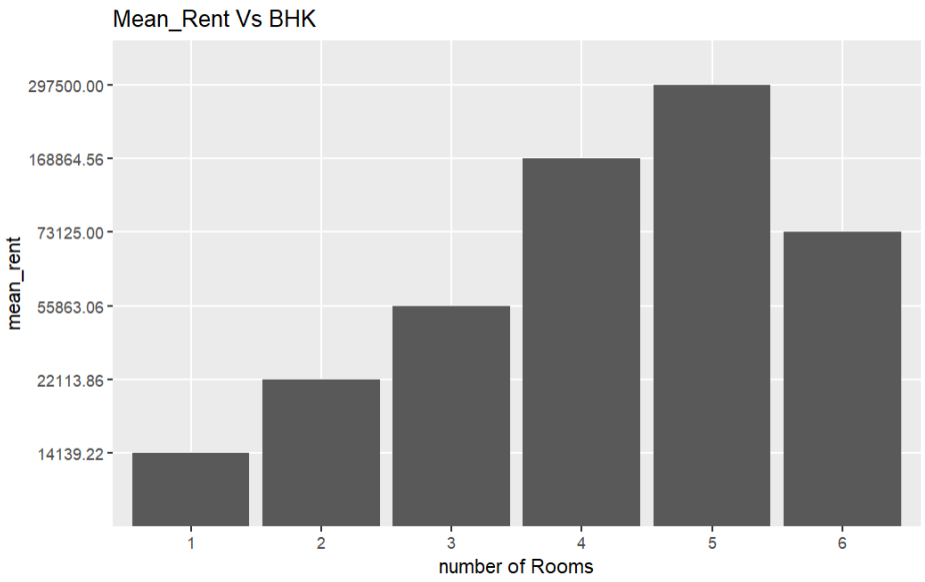
Variable explanation:

Column name	**Explanation**
BHK	Number of Bedrooms and Hall/Kitchen areas (BHK)
Rent	Rent of the properties
Size	Size of the properties in square feet
Floor	Floor of the properties and the total number of floors
Area Type	Size calculation method (super area, carpet area, or build area)
Area Location	Neighborhood location (Area Locality)
City	City where the properties are located
Furnishing	Furnishings status (furnished, semi-furnished, or unfurnished)
Tenant Preferred	Preferred type of tenant (Tenant Preferred)
Bathroom	Number of Bathrooms
Point of Contact	Point of contact for more information.

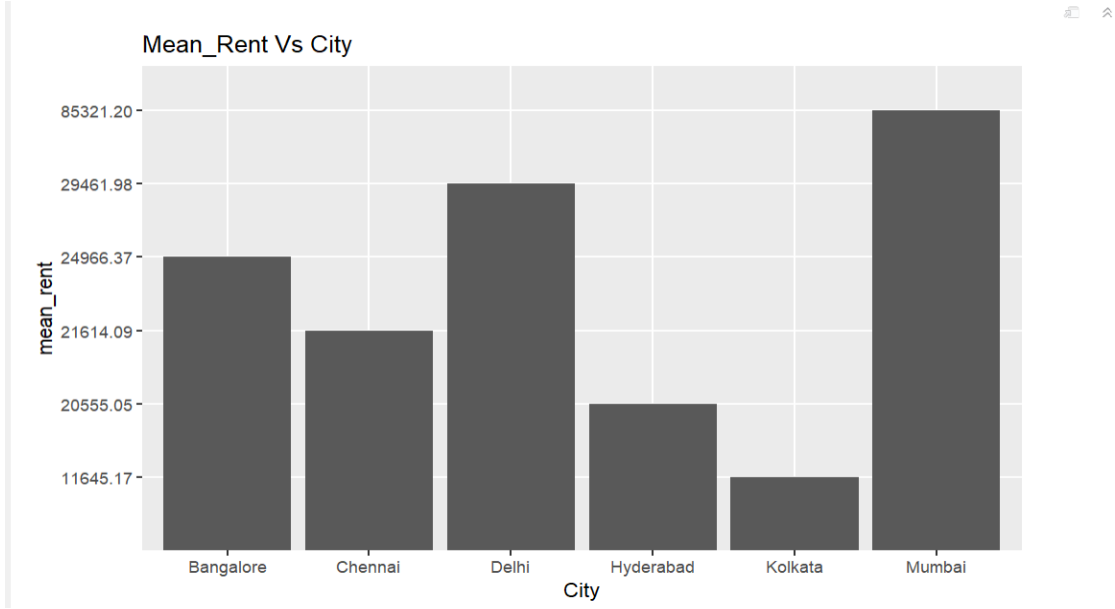
Therefore, the dataset has the size 4747*12 with 4747 rows and 12 columns. We can assume this data frame as a sample data frame for rental properties of the whole population in India. It has 12 variables where 5 of which are numerical variables and the rest of them are categorical variables. We will do various statistical analyses for these categorical and numerical variables from this 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.**

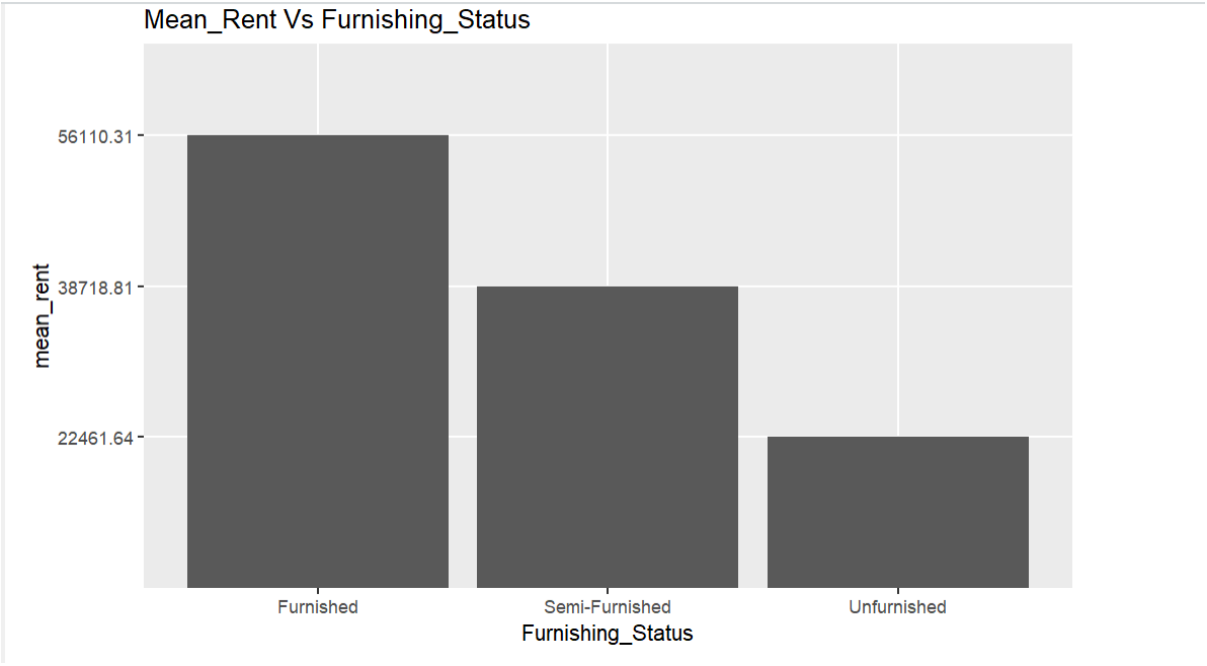
Answer: We will be using here **tydiverse package** for visualizing and analysing the dataset. This package includes a wide range of functions for cleaning and transforming data, as well as for creating informative and visually appealing visualizations. We will be using here one of the most popular features of tydiverse is the **ggplot package**, which provides an easy-to-use and highly customizable system for creating good-quality graphics.



From the above bar chart, we see that the mean rent price is highest for 5-room houses and lowest for 1-room houses.

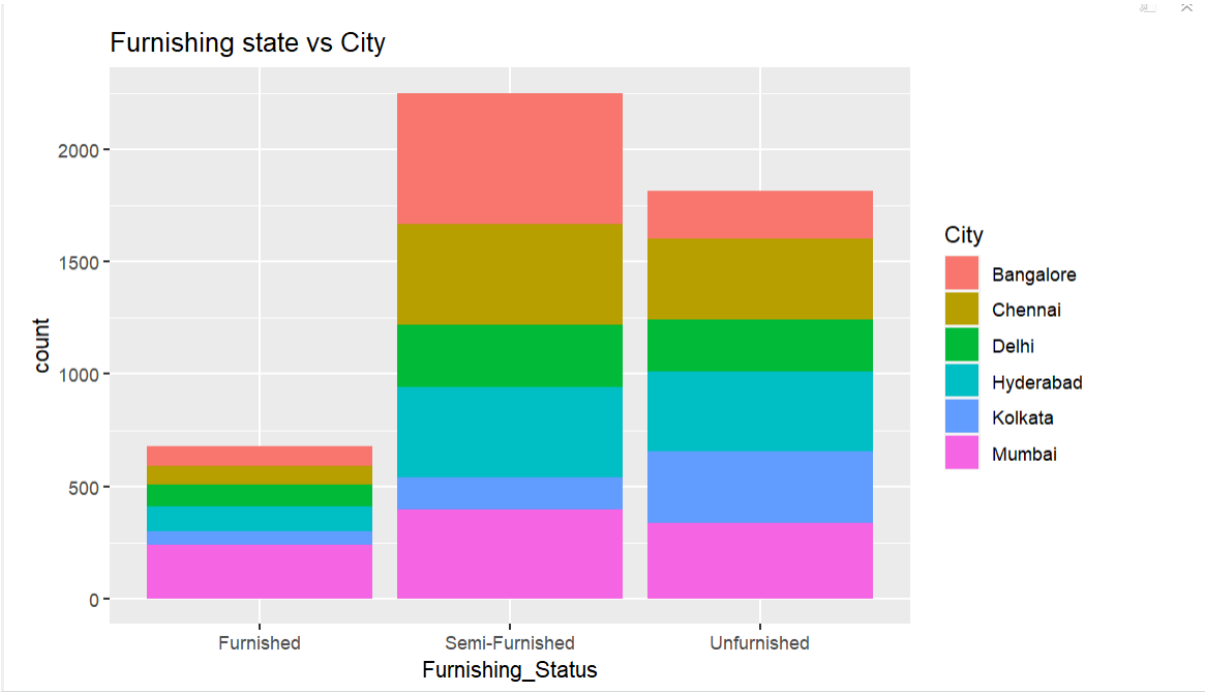


From the above bar chart, we see that the mean rent price is highest in Mumbai and lowest in Kolkata.



From the above bar chart, we see that the mean rent price is highest for Furnished houses and lowest for Unfurnished Houses.

2. **How is the furnishing state of the houses according to the cities?**
   Answer:



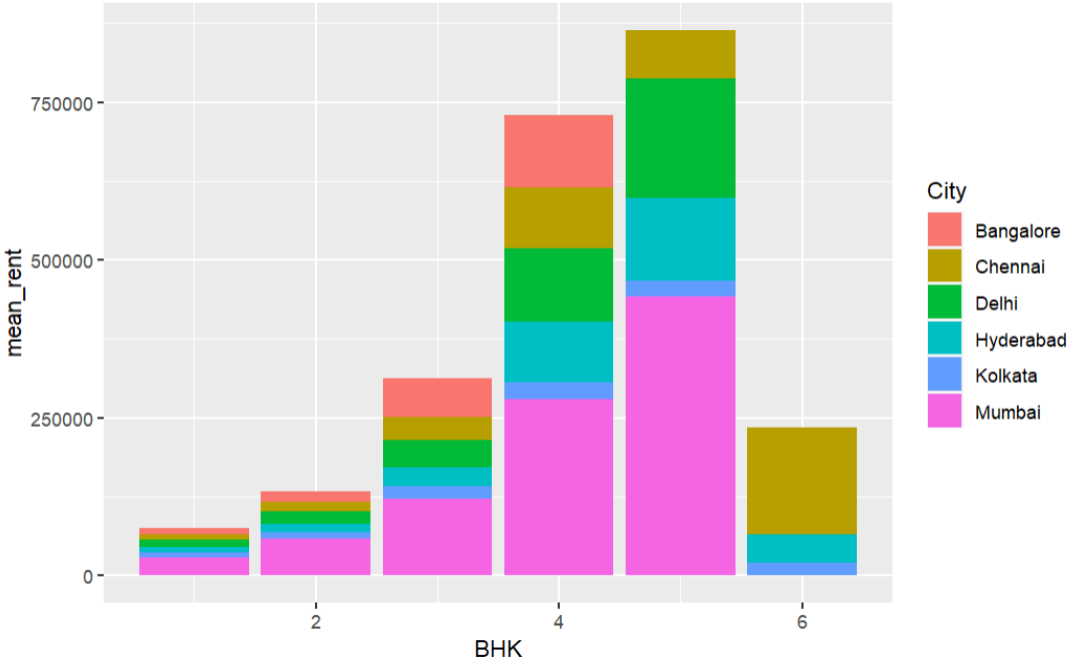
We see that, Semi-Furnished houses are mostly popular in India and Bangalore has the most Semi-Furnished houses where Mumbai has the most Furnished houses and Chennai has mostly Unfurnished houses among other cities.

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

Answer: We will be using **corrplot library** to create a correlation matrix using the "cor" function, and then visualize it using the "corrplot" function.

From the correlation Matrix, we see that Size has the highest correlation coefficient related to Rent compared to BHK. It means size has a fairly positive relationship with rent. Moreover, the Size and number of rooms(BHK) of the houses are also strongly correlated.

4. **How rent price varies with respect to the City and BHK?**
Answer:



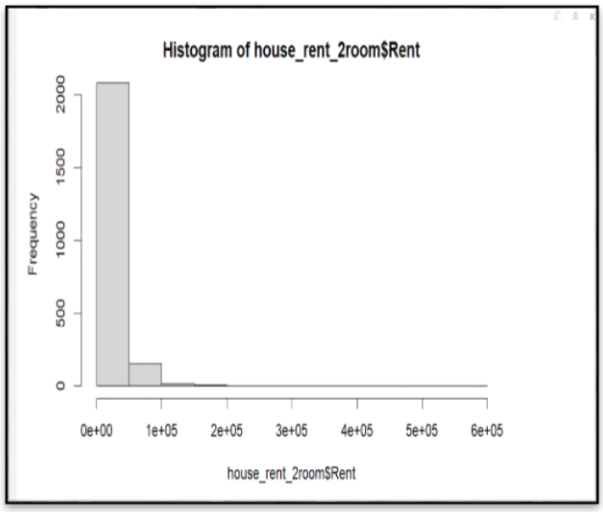
From the above graph, we can observe that Rent price is highest for all BHK types of houses in Mumbai. Only Chennai, Kolkata and Hyderabad are having 6-BHK houses with comparatively low prices.

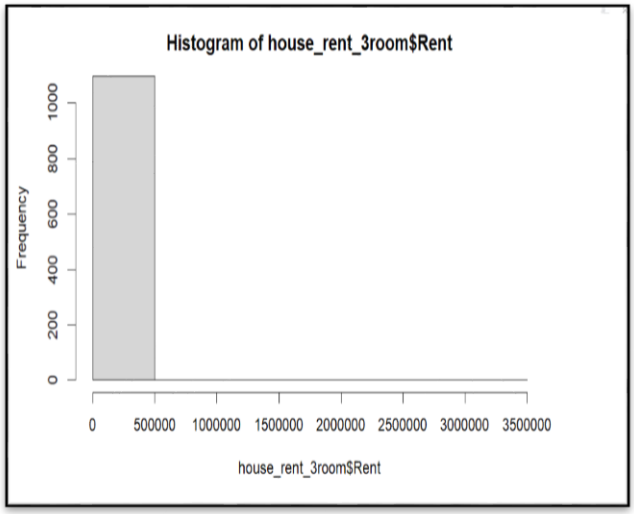
5. **From the given dataset, we will extract the Monthly rent data in Indian currency (Rupee) for 2-rooms and 3-room apartments in India and assume these as a sample to represent the population of all 2-room and 3-room house rent in India respectively. We will do multiple statistical analyses on this numerical data to conduct inference using bootstrap techniques and try to answer the following questions using R code**:

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

Ans:

To answer this, we need to plot the rent histogram and look at the distribution. If there are outliers present in the data, it is common to use the median instead of the mean because the outliers can greatly influence the mean, making it unrepresentative of the typical values in the data set.





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

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

Answer:

Bootstrapping: Resampling, or bootstrapping, involves randomly resampling a single dataset to create thousands of simulated datasets, producing meaningful results. The term "bootstrapping" refers to the impossible phrase of pulling oneself up by their own bootstraps. The bootstrap method has been around since 1979 and its usage has increased due to its ability to approximate correct sampling distributions. Bootstrapping does not create new data, but treats the original sample as a proxy for the real population and draws random samples from it. The resampling process creates many possible samples that a study could have drawn, allowing for the construction of confidence intervals and hypothesis testing. As the sample size increases, bootstrapping converges on the correct sampling distribution under most conditions.

Reasons for using bootstrapping:

1. Since our sample size is small compared to the whole population of India
2. Requires few assumptions to get a more accurate approximation

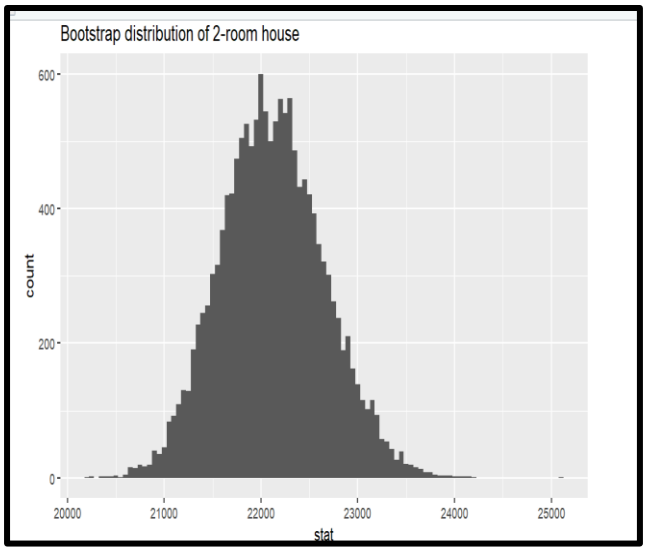
Using Bootstrapping in our problem: The bootstrap techniques can be used here to estimate the unknown population parameter (population mean of all 2-room and 3-room house rent in India) using only the given sample data. To do this will take random samples from the original sample with replacement multiple times to create a bootstrap population assuming it represents the whole population. Since we don't know the specific distribution, in this case, we will use non-parametric bootstrap with the following steps:

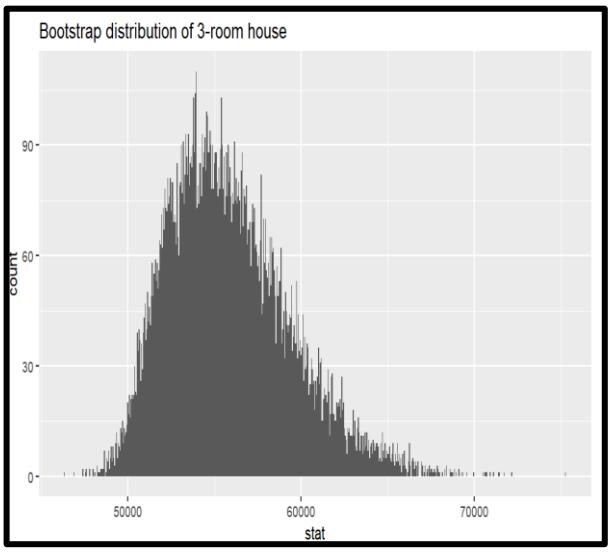
1. Defining the sample statistic of interest, in our case, it would be mean
2. Taking a random sample with a replacement of the same size as the original sample and repeating this process many times to obtain a series of bootstrap samples
3. Calculating the sample statistic for each resample/bootstrap
4. Getting an estimate of the bootstrap statistic from the bootstrap distribution.
5. Estimating confidence intervals for the population statistic, for example, by computing the 2.5th and 97.5th percentiles of the distribution of resample/ bootstrap statistics.

Bootstrapping using Infer Package:

Infer Package: Infer is a package within the **tydiverse package** for statistical inference. It provides key verbs such as specify(), hypothesize(), generate(), and calculate(). These verbs c specify the variable of interest, create the null hypothesis, generate realizations, and calculate test statistic. It's designed to work with tibbles and multiple columns of data. Infer is useful for simulation-based tests and calculating appropriate test statistics for each sample.

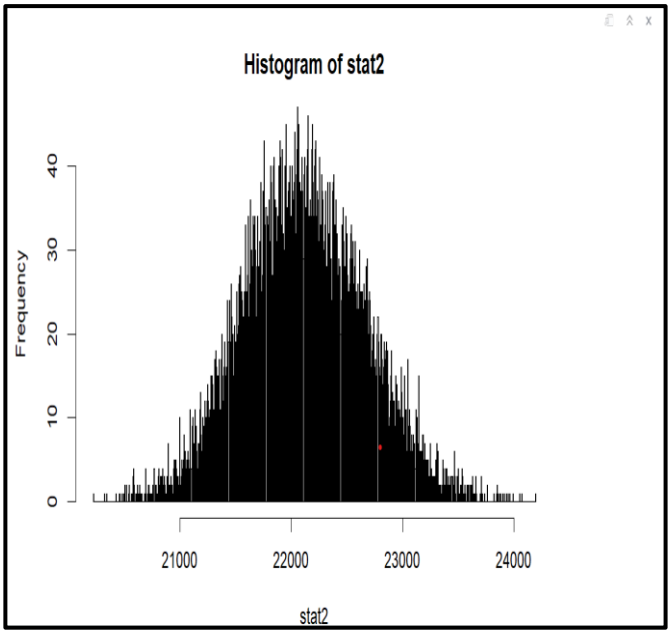
The following figures are the bootstrap distributions comprising 10000 bootstrap means for both 2-room and 3-room house rents.

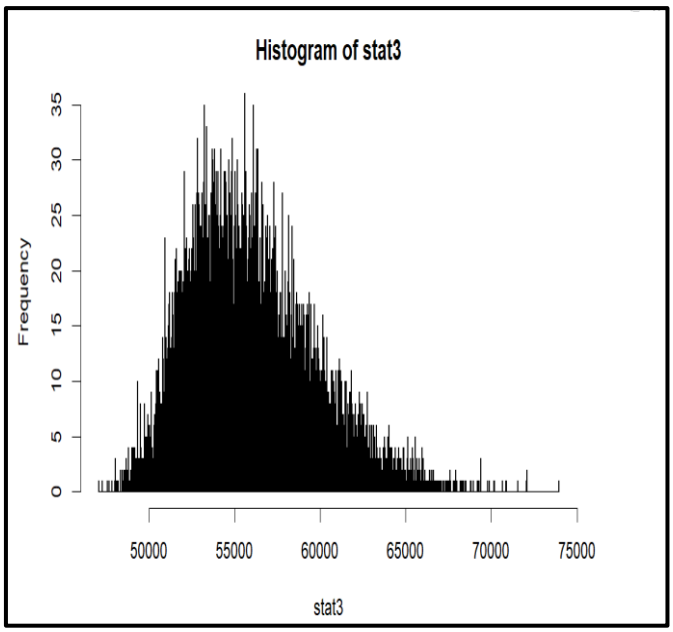




Bootstrapping manually:

The following figures are the bootstrap distributions, comprised of 10000 bootstrap means for both 2-room and 3-room house rents.



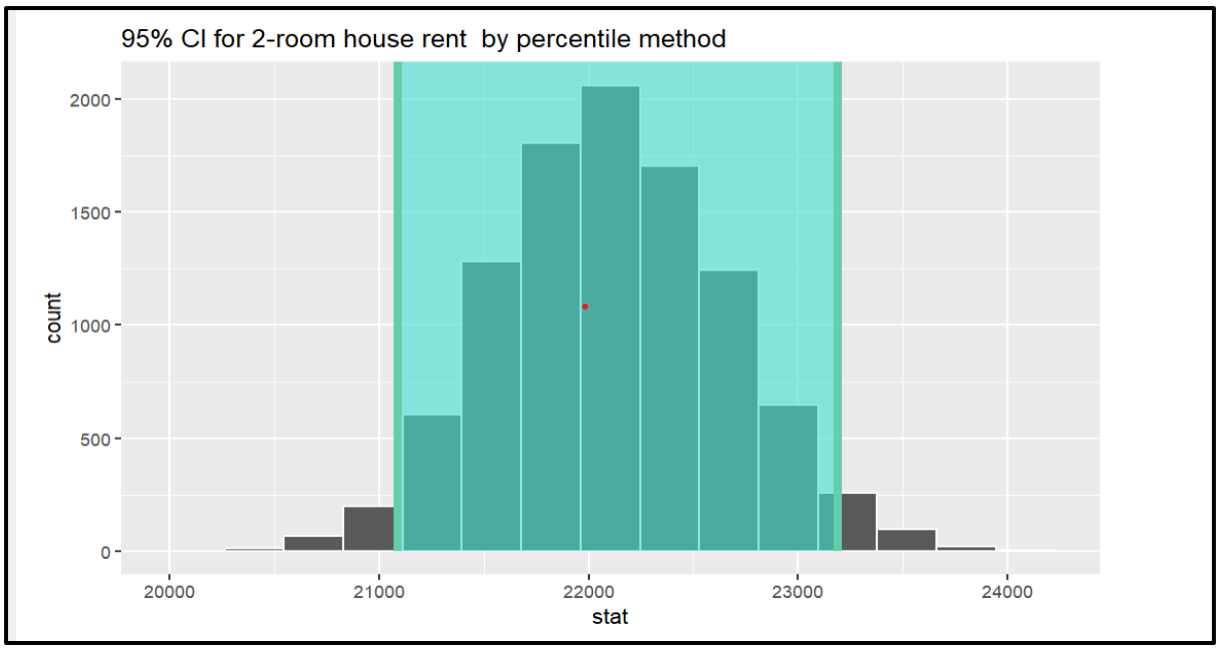


We get the same distribution figures for both manually and using packages.

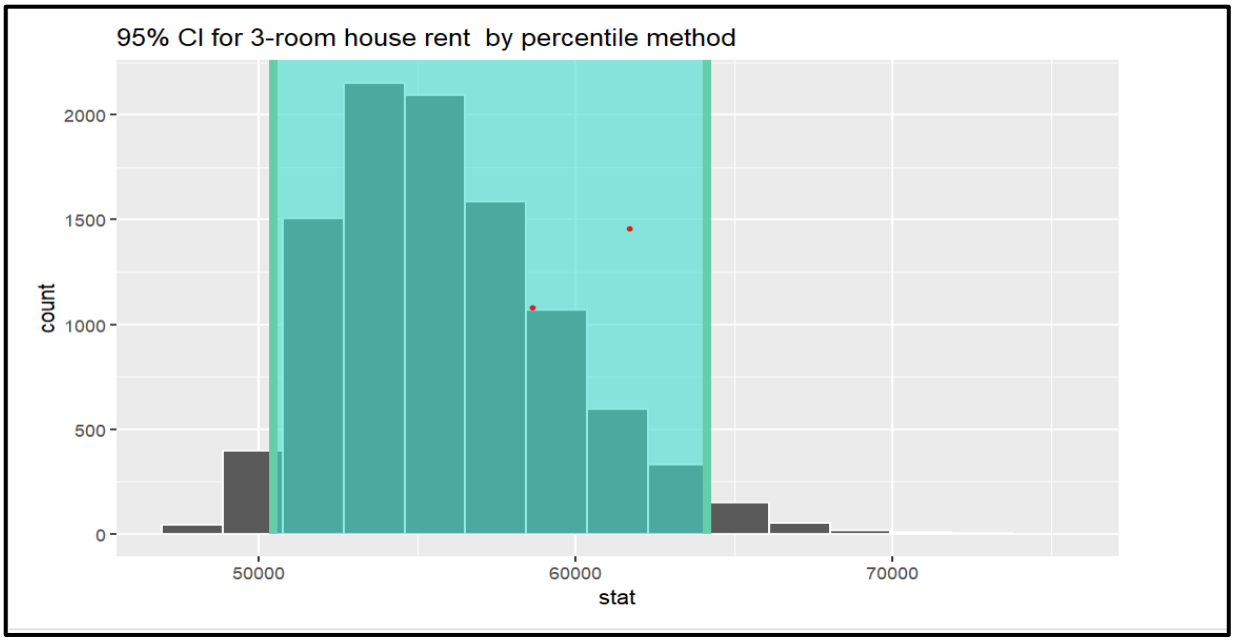
Calculating Confidence Interval:

Using the obtained non-parametric bootstrap distribution, we can calculate confidence intervals for the mean price of 2-room and 3-room houses in the following ways.

1. **Percentile method:** we can estimate the 95% percentile by CI using the boundary of the middle 95% of the bootstrap distribution.
   Result from Percentile Method: we get (21090.57, 23188.7) as the 95% CI for the mean rent of all 2-room houses in India. That is, we are 95% confident that the mean rent of all 2-room houses is between 21090.57 Rupee and 23188.7 Rupee.



we get (50480.31,64148.54) as the 95% CI for the mean rent of all 3-room houses in India. That is, we are 95% confident that the mean rent of all 3-room houses is between 50480.31 Rupee and 64148.54 Rupee.



2. **Bootstrap-t method :** we can estimate the 95% percentile by estimating CI using the formula,

$$[sample\ statistic + t_{\frac{\alpha}{2}} * SE(sample\ statistic), \hat{\theta} + t_{1-\frac{\alpha}{2}} * SE(sample\ statistic)]$$

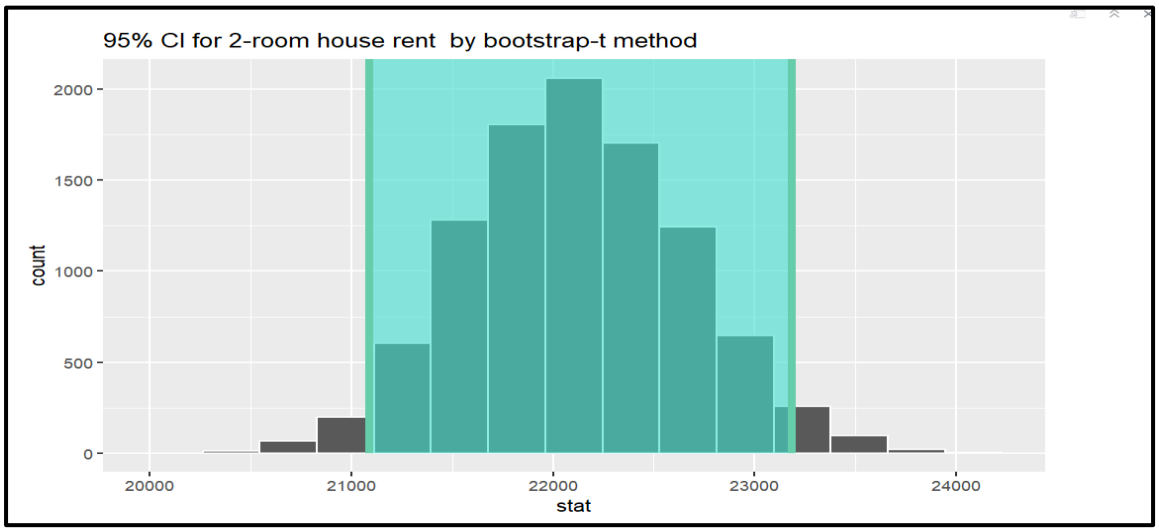
Where, SE is the standard error and bootstrap t-values are calculated by

$$t_b = \frac{bootstrap\ statistic - sample\ statistic}{bootstrap\ standard\ error} \quad for\ b = 1, \dots, B$$

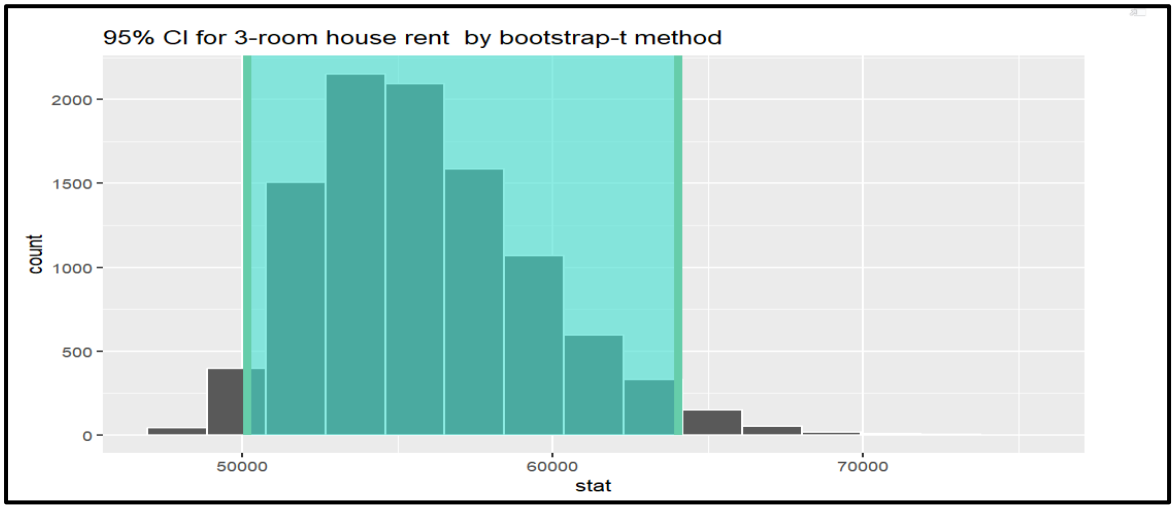
Here, B is the number of independent bootstrap samples taken from the original sample
Then we can calculate $t_{\frac{\alpha}{2}}$ and $t_{1-\frac{\alpha}{2}}$ by quantiles of t_b.

Results from Bootstrap-t Method:

we get (21086.15, 23187.63) as the 95% CI for the mean rent of all 2-room houses in India. That is, we are 95% confident that the mean rent of all 2-room houses is between 21086.15 Rupee and 23187.63 Rupee.



we get(50186.62,64056.98)as the 95% CI for the mean rent of all 3-room houses in India. That is, we are 95% confident that the mean rent of all 3-room houses is between 50186.62 Rupee and 64056.98 Rupee.



We see that we got similar result from both percentile and bootstrap-t method.

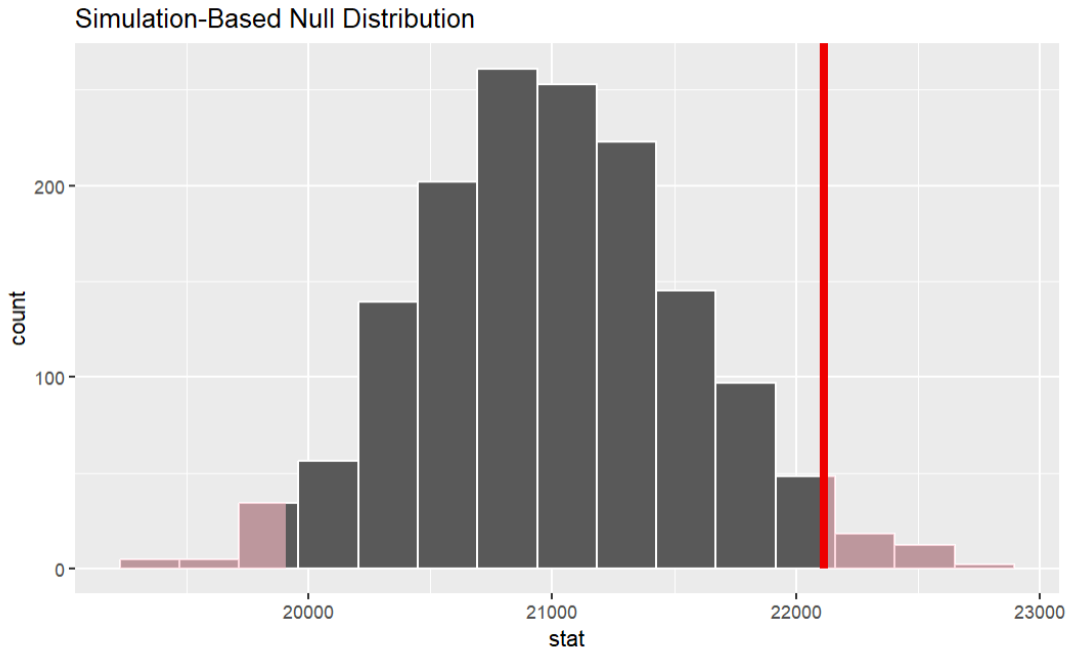
iii. **Evaluate whether this data provides evidence that the mean rent of 2-room houses in India is different than 21000 Rupee?**

Answer: we can use bootstrapping here to do the point null-hypothesis test here. Here, we assume the null hypothesis as, H_o: mu=24000 and alternative hypothesis as H_1: mu != 21000

Hypothesis test: It is used to determine whether there is sufficient evidence to reject a null hypothesis in favor of an alternative hypothesis. The 5% significance level, also known as the alpha level, is a commonly used threshold for determining statistical significance. If the p-value (the probability of obtaining a test statistic as extreme or more extreme than the one observed under the null hypothesis) is less than or equal to 0.05, then the result is considered statistically significant, and the null hypothesis is rejected in favor of the alternative hypothesis.

Point null hypothesis testing using bootstrapping: Point null hypothesis testing is a statistical hypothesis testing approach where a null hypothesis is formulated as a point value for a parameter of interest. For example, in the case of a mean, the null hypothesis might be that the mean is equal to a specific value. In point null hypothesis testing using bootstrapping, the null hypothesis is tested by comparing the observed value of the statistic to its estimated sampling distribution obtained through bootstrapping.

Result from R code: We got the p-value is 0.048 , which is below the significance level, so we can reject the null hypothesis in favor of the alternative hypothesis, and conclude that the data provides convincing evidence for the alternative hypothesis.



Conclusion: In conclusion, the dataset of 4747 rental properties in India was analyzed to gain insights into the rental property market in India. The dataset had 12 variables, including 5 numerical variables and 7 categorical variables, and was used to conduct various statistical analyses such as correlation analysis, confidence interval and hypothesis testing using bootstrap techniques, and visualization using the tydiverse package. The analyses revealed that the rent price varies significantly based on the number of rooms, city, and furnishing status. The semi-furnished houses were found to be the most

popular in India, and Mumbai had the highest average rent price. The bootstrap technique was used to estimate the population mean of rental prices for 2-room and 3-room houses in India. Overall, the study provides useful information for individuals and organizations interested in the rental property market in India.

Appendix: The corresponding R code has been attached as Assignment-4.Rmd file.

Reference:

1. https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset?select=House_Rent_Dataset.csv
2. https://statisticsbyjim.com/hypothesis-testing/bootstrapping/
3. https://yards.albert-rapp.de/hypotests#bootstrap
4. https://jrnold.github.io/r4ds-exercise-solutions/exploratory-data-analysis.html
5. Non-Parametric Bootstrap-Data visualization and Modelling- Pere Puig