**Housing Solutions for Northeastern Students**

By

Aanchal Chholak

Table of contents

[1. Introduction 2](#_Toc11275)

[2. Data collection 2](#_Toc11276)

[3. Data visualization 4](#_Toc11277)

[a. Histogram 5](#_Toc11278)

[b. Bar Graph 6](#_Toc11279)

[c. Scatter Plot 7](#_Toc11280)

[4. Statistical analysis 8](#_Toc11281)

[a. Sampling distribution of mean and variance 8](#_Toc11282)

[b. Hypothesis testing 9](#_Toc11283)

[c. Confidence Intervals 10](#_Toc11284)

[5. Conclusion 12](#_Toc11285)

[6. References 12](#_Toc11286)

# 1.Introduction

The problem faced by most students while relocating to Boston is finding housing. As there are a lot of factors to be considered while searching for accommodation, students tend to fall for frauds due to their lack of knowledge about the city. The plan of action is to perform a descriptive analysis of the dataset which will enable the student to make an informed decision when it comes to selecting the ideal home. The scope of the project is conducting statistical analysis to study various characteristics such as distance from the university, public transport options, locality, rent, etc. based on which student selects housing. The motive will be to find ideal accommodation based on the characteristics and data that the project concludes. Furthermore, determining the deviation from the ideal characteristics will help the student to differentiate and choose between the available options according to his/her liking.

# 2.Data collection

The data for visualization and statistical analysis was collected by conducting a survey. With the help of the survey, the data like, total rent, individual rent, maximum budget, number of rooms, number of roommates, use of public transport, cost of public transport, location of accommodation and preferred amenities were collected. The collected data was then customized in excel into 2 types for ease of visualization.

A screenshot of a computer

Description automatically generated

Table 1: Sample of data

The table above (see Table 1) is a sample of the data collected from the survey. Given below are the variables for each column.

1. totalrent – Total rent of the accommodation.
2. individualrent – Individual rent paid by the student in that accommodation.
3. maxbudget – The maximum rent that a student can afford.
4. rooms – Total number of rooms in the accommodation.
5. roommates - Total number of roommates in the accommodation. 6. publictransport – If a student uses public transport or not
6. cost – Transport cost paid by the student.
7. location – The location of the accommodation. It was divided into 9 different regions viz. Allston, Brighton, Brookline, Cambridge, Fenway – Kenmore, Jamaica Plain, Mission Hill, Roxbury and Other.
8. conveniencestore – Convenience store accessibility as a preferred amenity of a student.
9. pharmacy – Pharmacy accessibility as a preferred amenity of a student.
10. gymnasium – Gymnasium accessibility as a preferred amenity of a student.
11. transport – Public Transport accessibility as a preferred amenity of a student.

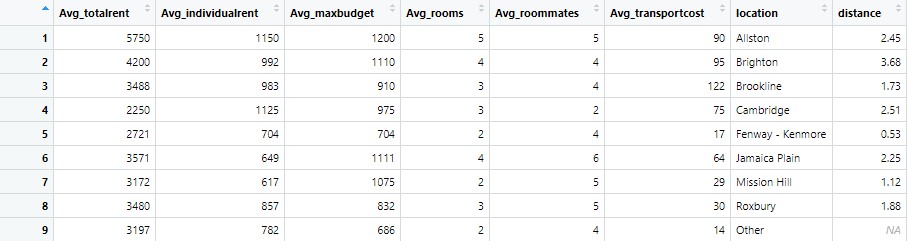


Table 2: Data for bar graph

The table above (see Table 2) is summarized for visualization using Bar graphs. In this table the average of all the values like total rent, individual rent, maximum budget, number of rooms, number of roommates, cost of public transport was calculated based on the Location. Given below are the variables for each column.

1. Avg\_totalrent – Average of total rent based on the Location.
2. Avg\_individualrent - Average of individual rent based on the Location.
3. Avg\_maxbudget - Average of maximum budget based on the Location.
4. Avg\_rooms - Average of number of rooms in the accommodation based on the Location.
5. Avg\_roommates - Average of number of roommates in the accommodation based on the Location.
6. Avg\_transportcost - Average of cost of transport based on the Location.
7. distance – Distance of the accommodation location from the University.

# 3.Data visualization

Chart, scatter chart

Description automatically generated

Figure 1: Quartile – Quartile Plot

The quartile - quartile plot of the total rent was obtained to verify the normality of the data. In the given graph (see Fig. 1), the diagonal line represents the normal distribution. The points describe the distribution of total rent. The points close to the diagonal lines signify that the data is normally distributed.

## a.Histogram

Chart, histogram

Description automatically generated

Figure 2: Histogram of Total rent vs Frequency & Density curve

The above graph was formed from the ‘total rent’ values. In the graph, a histogram as well as the density curve was plotted. From the histogram, it can be observed that most of the total rent paid by the students is below 4000. Also, with the approximate bell shape of the density curve, the total rent data is normally distributed.

Chart, bar chart

Description automatically generated

Figure 3: Histogram of Location vs Frequency

The above graph was obtained from the Locations mentioned by the student vs the frequency of the accommodation location. From the graph, most of the students are living the Roxbury area.

## b.Bar Graph

Chart, bar chart

Description automatically generated

Figure 4: Bar graph of Location vs Average individual rent

The above bar graph of Location vs Average individual rent was obtained from the data of the summarized table (see Table 2). In this graph, the average cost paid a student in an accommodation in a particular location was plotted. It was observed that the lowest average individual rent paid by a student was in Mission Hill location. While the highest average individual rent paid was in Allston area.

Chart, bar chart

Description automatically generated

Figure 5: Bar graph of Location vs Average transport cost

The bar graph (see Fig. 5) of Location vs Average transport cost was obtained from the data. In this graph, the average transport cost paid by a student to commute to the university for their accommodation location was observed. Average transport cost paid by a student in Brookline area was the highest while it was lowest for a student in Fenway – Kenmore area.

## c.Scatter Plot

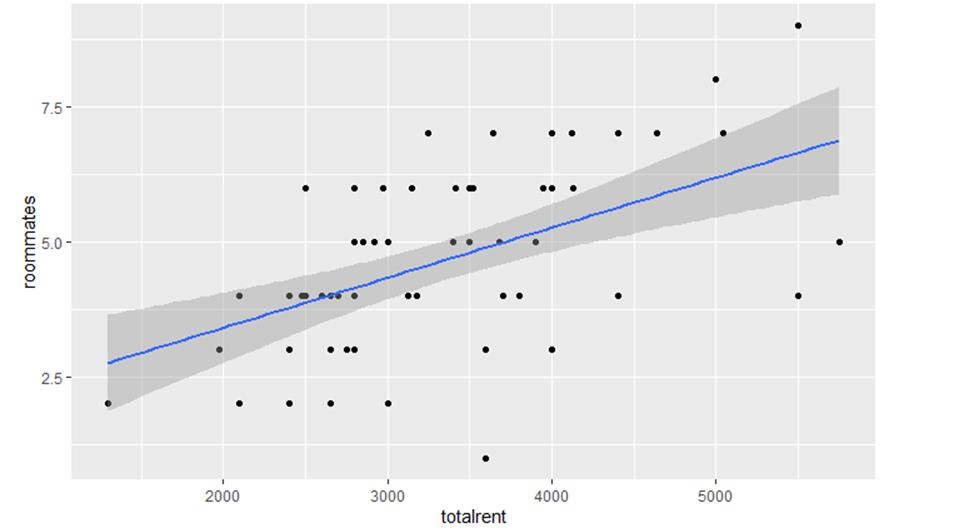


Figure 6: Scatter plot for Total rent vs Roommates

In the above graph it is observed that the scatter plot for Total rent vs roommates shows a positive correlation. To simplify it further, the above graph concludes that as the total rent for an accommodation increases the number of roommates within a single household increases. The reason behind this is, to tackle the overall high price, students usually rent out a single accommodation in groups where a single group on an average can consist of 4 to 5 members. This helps to reduce the rent that is to be paid by a single individual.

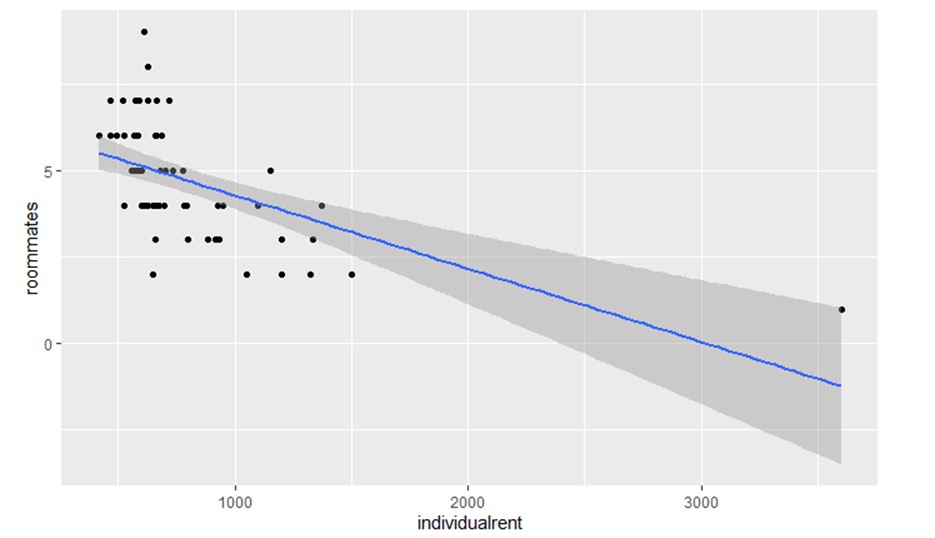


Figure 7: Scatter plot for Individual rent vs Roommates

The above scatter plot for Individual rent vs roommates shows a negative correlation which means as the amount for individual rent increases the number of roommates decreases. To reason behind this is that certain individuals are ready to pay a premium in terms of rent for private rooms and in some cases entire accommodations are rented out by a smaller group of 2 to 3 members to better utilize the available resources.

# 4.Statistical analysis

## a.Sampling distribution of mean and variance

The sampling distribution for the three main characteristics were found using R. i.e., Total rent, Individual rent, Cost of transportation, Distance from the university. Firstly, a sampling distribution was generated. The following code was used to generate a random sampling.

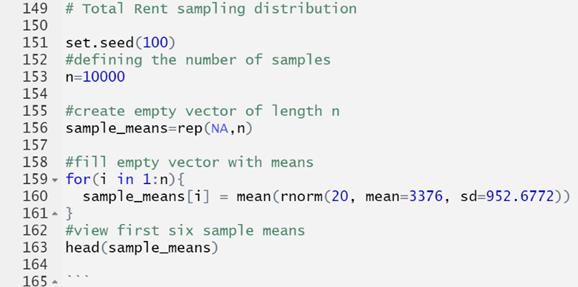


Figure 8: R code for random sample

In this example, rnorm() function was used to calculate the mean of 10,000 samples in which each sample size was 20 and was generated from a normal distribution with a mean of 5.3 and standard deviation of 9. The first sample had a mean of 3478.763, the second sample had a mean of 3464.110.

Next, a Histogram was created to visualize and to check whether the sampling distribution is bell shaped with a peak of 200. Then the mean and standard deviation of the sampling distribution was calculated. It was observed that the actual sampling mean is very similar to the theoretical mean of the sampling distribution. Same for the standard deviation.

Given below is the table for the theoretical mean & standard deviation and actual sampling mean & standard deviation.

|  |  |  |
| --- | --- | --- |
|  | Theoretical Mean | Actual Sampling Mean |
| Total Rent | 3376 | 3379.11 |
| Individual Rent | 809.8374 | 811.2988 |
| Distance from Uni | 1.456 | 1.46 |
| Amenities | 1.9666 | 1.969286 |

Table 3: Theoretical and Actual Mean and Standard Deviation

The following observations were obtained from the probabilities found for the important characteristics.

1. If the total rent of a house is considered, there is a scarcity of houses with rent less than 3300 (with mean being 3380) at 35% and a large selection of houses with rent less than 3500 at 71%.
2. For individual rent of a house, considering the mean as 809, there is 81% chance to find housing less than 900 and there is a 59% chance of it being more than 700.
3. The probability of finding a house within 2 miles from the university is 98.4%. which means that at the above price point there is pretty good chance of finding a house
4. From the analysis It can be observed that a student requires at least 2 amenities from the selection, and the probability of them needing just one is 56%.

## b.Hypothesis testing

A survey was conducted for 60 students of northeastern university having average rent 809.83 and standard deviation 447.69 as calculated from population sample. Can it be regarded as the ideal individual rent drawn from the normal distribution with assumed mean having 95% confidence intervals.

Case 1 – 700

**Step-1 Determine the value for null and alternative hypothesis**

* Ho = 700
* Ha ≠ 700

Assuming ideal rent = 700$,

**Step-2 Applying the hypothesis testing and Z Test (since n>=60)**

A picture containing logo

Description automatically generated

The z value after calculating the above formula is 1.90

**Step-3 Comparing the z value at 95% confidence level**

The Z value from Z table for 95% confidence level is 1.96 which is more than statistics value 1.90. The population mean value does not lie in the rejection region as shown in graph (see Fig. 9), hence it can be concluded that the null hypothesis is accepted.

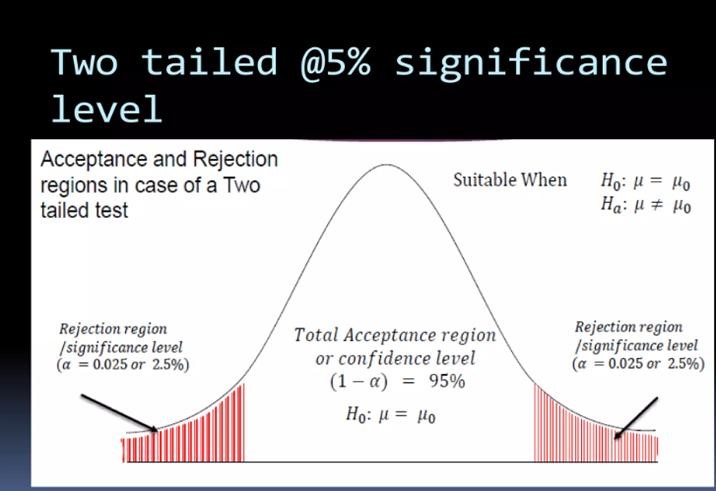
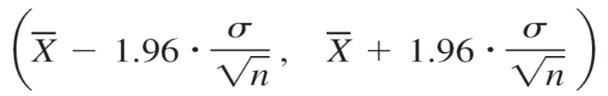


Figure 9: Two tailed test graph

## c.Confidence Intervals

Considering sample size, N = 60, it was found that CI for the nearest housing locations. The area under the standard normal curve lies between -1.96 and +1.96 for confidence interval taken at confidence level 0.95.

CI is given by the following formula: -



From the data collected in the survey 3 main questions are answered: -

1. **What is the most accessible location for the students from the University?**

Sample size (n) = 60

Sample mean (x) = 1.812

Std deviation(s) = 0.78

For this problem, the Lower value is 1.615 miles and Upper Value is 2.009 miles from the University.

1. **What is the most reasonable rent for the students?**

Sample size (n) = 60

Sample mean (x) = 809.9

Std deviation (s) = 447.67

The intervals obtained are having the Lower value as $697 and Upper Value as $923.

1. **What is the most affordable transport cost for students?**

Sample size (n) = 60

Sample mean (x) = 56.33

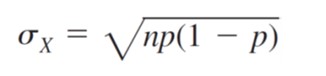
Std deviation (s) = 50.52

The intervals obtained are having the Lower value as $44 and Upper Value as $69.

**Confidence Interval for Population Proportion**

To find the confidence interval for a population proportion, following analogy has been used – X = number of successes in the sample p = proportion of success in a population

As the population size is much bigger that the sample size, it is considered that X is a binomial random variable where E(X) = np and the standard deviation is given by –



The point estimate for p is given by p̂: - 

1. For finding the most accessible preferred location to the University, here, p̂ is the natural estimator of p which is defined as the sample fraction of the people who live within 1.615 and 2.09 the miles distance from the University.

These are the confidence intervals obtained for the proportion of distance – n = 60, X = 23

Lower interval – 25.7%

Upper interval – 50.3%

1. For finding the most reasonable individual rent, here, p̂ is the natural estimator of p which is defined as the sample fraction of the people who have found a housing with the average affordable individual rent.

These are the confidence intervals obtained for the proportion of individual rent – n = 60, X = 13

Lower interval – 11.5%

Upper interval – 32.5%

1. For finding the most reasonable individual rent, here, p̂ is the natural estimator of p which is defined as the sample fraction of the people who pay between $44 and $69 for monthly commute.

n = 60, X = 11

Lower interval – 8.3%

Upper interval – 27.7%

# 5.Conclusion

The goal of this project was to analyze the characteristics of housing data from the survey that was collected, to predict the ideal characteristics. The first step was undertaking the descriptive statistics to obtain the shape and distribution of the data and check its normality wherein the histogram and the density curve gave the majority of the total rent paid by the student peaking at around 3000usd and lying below 4000usd. Following sample data statistical analysis, further get population distribution and hypothesis testing. Probability of individual rent, total rent of house, Distance from university when there is an event of random samples selected from the population was found. Later performing the sampling distribution of means and variance that there is a 59% chance of finding a house with rent more than 700 and less than 900. After hypothesis testing it was found that for individual rent 700 lies between the ideal range.

Considering the intervals, it is inferred from the data that the most accessible locations from the University are Longwood Medical Area, Brookline, Fenway-Kenmore, and Mission Hill. The most reasonable rent for students is approximately around $800. The students can afford $60 for monthly commute. In conclusion, the most accessible, reasonable area is Mission Main with rent around $720 and bus being the most affordable mode of transportation at $55 per month.

# 6.References

* “Probability and Statistics for Engineers & Scientists” – Ninth Edition by R. E. Walpole, R. H. Myers, S. L. Myers, K. Ye
* [https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-ztest.html](https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-z-test.html)
* <https://rpubs.com/pythonjokeun/samplingdistribution>
* [http://www.sthda.com](http://www.sthda.com/)
* <https://www.simplypsychology.org/confidence-interval.html>