ALY 6010 Final Report

Ifeoluwa Adegbohungbe

College of Professional Studies, Northeastern University

Dr. Andy Chen

16th December 2023

Introduction

The exploratory analysis report for this project was conducted on a dataset extracted from a Portuguese real estate website by Marco Carujo on the 21st of March, 2021. The data contained information on properties to be rented or sold, their size, location, price, condition and the number of rooms/bathrooms they have. Using the tidyverse package, this data was cleaned by removing extraneous columns and formatting the data. While making the first few graphs, NAs and outliers were removed and new columns were created as needed. Upon cleaning and running several plots, it was discovered that the cities of Faro, Lisboa, Setúbal, Porto, Braga and Erova have a very small portion of their advertised properties up for rent and Faro is the only city that has both larger and more expensive houses. I came to the conclusion that properties in the expensive areas of Portugal are unlikely to be rented (see Figure 3) and that although there might be a correlation between property prices and their size (according to Figures 1 and 2), it is not very strong.

Although the correlations between price and other variables can vaguely be described by the visualisations created in earlier reports (Figures 1 and 2), this method can only be applied to about ten cities at a time. This also requires looking at two plots simultaneously as the difference in magnitudes make it difficult to see any trends when the variables are plotted on the same graph (Figure 4). In this report, I want to quantify the correlation between price and size across all recorded cities.

With Pearson's product-moment correlation function, the comparison can consider all data points for price and area. This test has a very low p-value for the null hypothesis that the price and area have no correlation. This is in line with what is shown on the graphs that sample the most expensive cities (Figures 1 and 2). The alternative hypothesis of this test would be that

the correlation coefficient between the variables is not equal to zero. This test also gives an estimate of the correlation coefficient, 0.052. This means although the price and area of the properties share a positive correlation, it is a very weak one.

During the statistical analysis of this dataset, I concluded that the average price of property in Portugal is within €472,833 and €482,735 using the two-sample T-test for means. This test was done by creating two sampling distributions with 2,000 samples each (see Figures 5 and 6). As both were normally distributed, they fit the criteria to use this test. Most of the sample prices were between €450,000 and €500,000 for both distributions. Due to this similarity, the first test, with a null hypothesis that the means were equal, was conducted with a confidence level of 99%. The result was a p-value of 0.196, signifying that the difference between the means, €173.2 was statistically significant. As a result, the alternative hypothesis, that the means were not equal, was proven to be true. Next, I tested for the true mean of the price column being €478,550. This null hypothesis was constructed by selecting a value between €150 of the means of the sampling distributions. This test yielded a p-value of 0.799, meaning that the true mean falls within the 90% confidence interval of €472,833 to €482,735.

In this report, I will test for the statistical power of this conclusion. Since the confidence level used was 90%, the significance level is 0.1. Since the distributions have the same sample size, the function pwr.t.test() is used. However, the function cannot provide the effect size since the power is also unknown. The effect size is calculated with the differences in the samples' means and standard deviations using esc_mean_sd() from the esc() package. The statistical power of the conclusion is 0.364, which is not very high, despite the distributions being so similar. I believe this may be due to the number of samples used. A test shows that running the test with distributions containing 5,000 samples would double the power of the conclusion.

Conclusion

In this report, I have used hypothesis tests to back up the conclusions that were previously made by looking at graphs. The numeric conclusions are much better as their certainty can be qualified. The conclusion of the average prices is less reliable than the correlation between price and area, of which we can be very certain.

Works Cited

Bluman, A. G. (2012). *Elementary statistics. A step by Step Approach* (8th ed.). McGraw-Hill.

Carujo, M. (2023, June 17). *Portugal properties - rent, buy and vacation.* Kaggle. https://www.kaggle.com/datasets/mcarujo/portugal-proprieties-rent-buy-and-vacation

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. Boca Raton, FL and London: Chapmann & Hall/CRC Press. ISBN 978-0-367-61007-4.

Appendix

Figure 1. Average prices per city of the cities with the most expensive properties

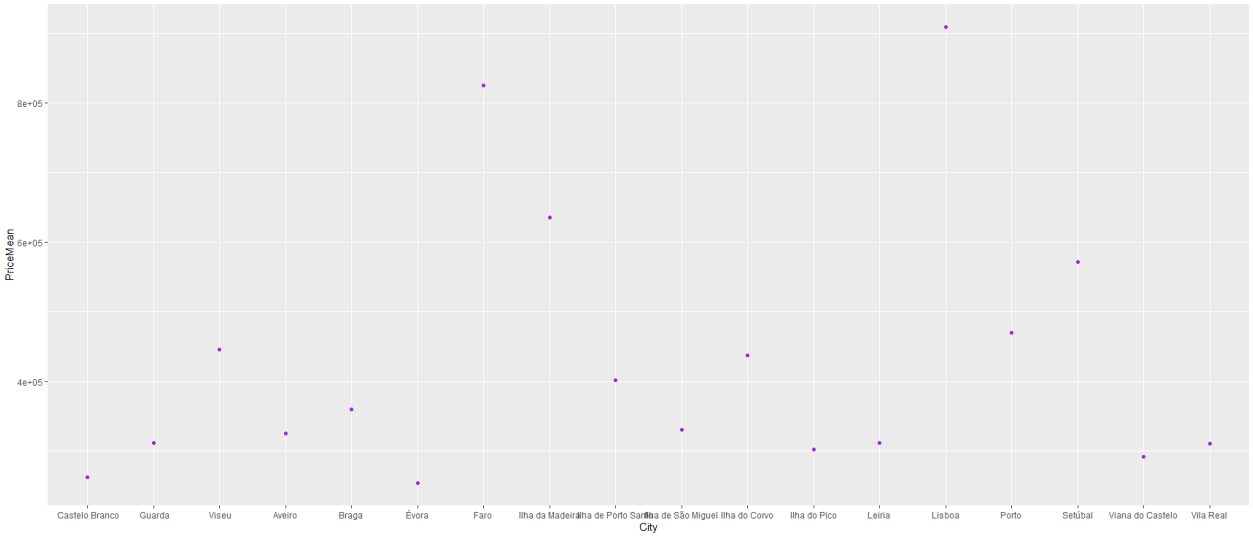

Figure 2. Average area per city of the cities with the largest properties

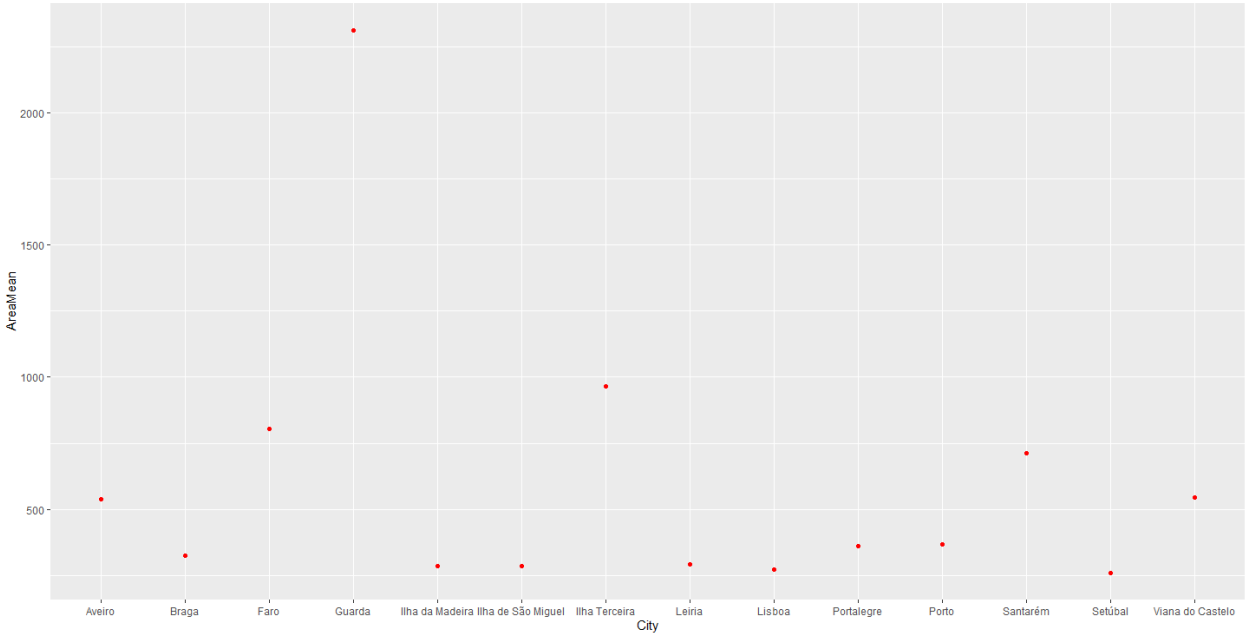

Figure 3. Distribution of rent, sell and vacation among the cities with the most expensive properties

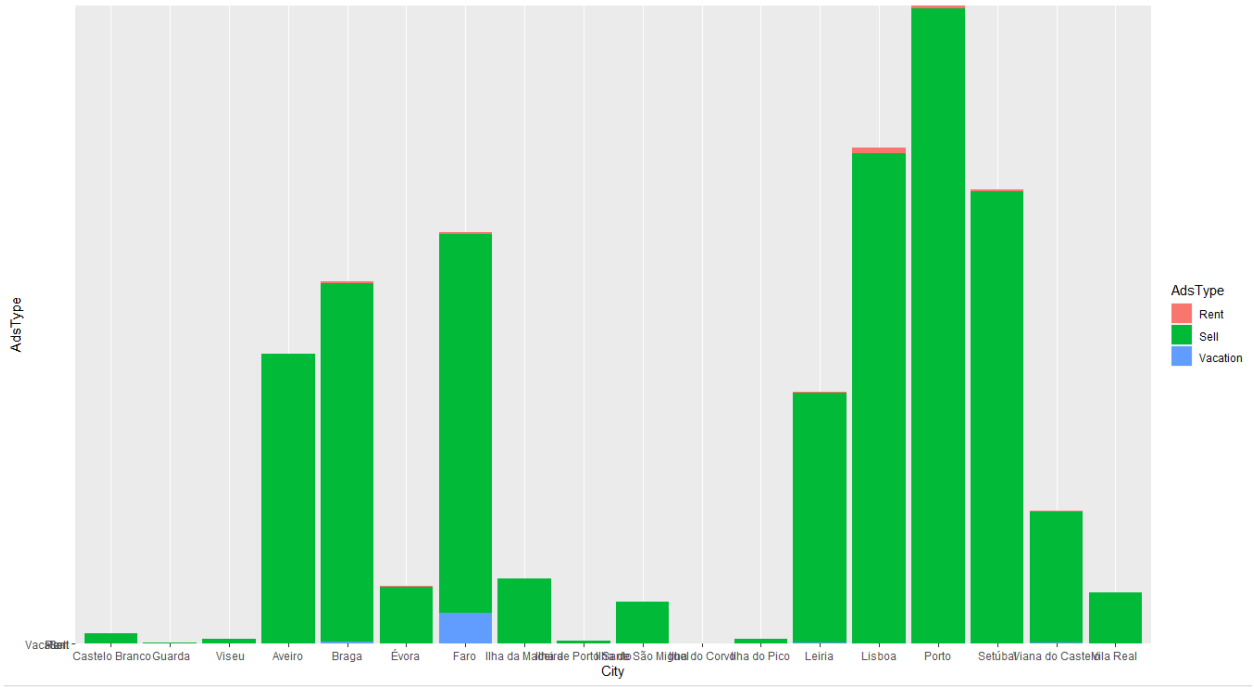

Figure 4. Magnitude difference between Price and Area prevents direct comparison

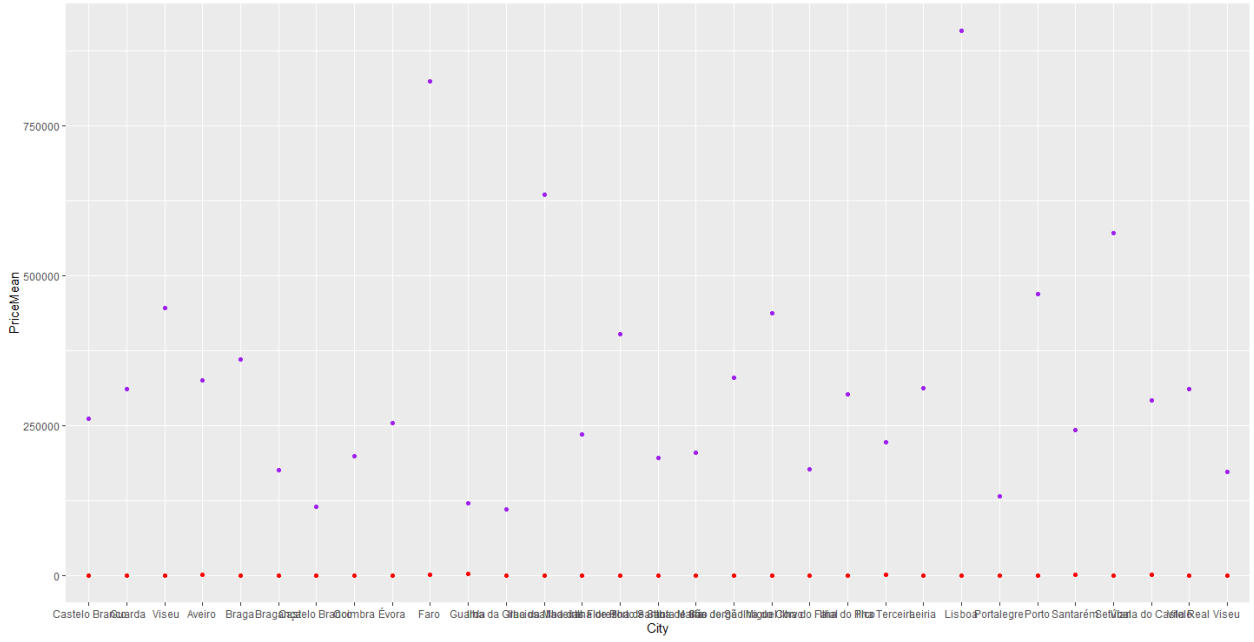

Figure 5. Histogram of the first sampling distribution

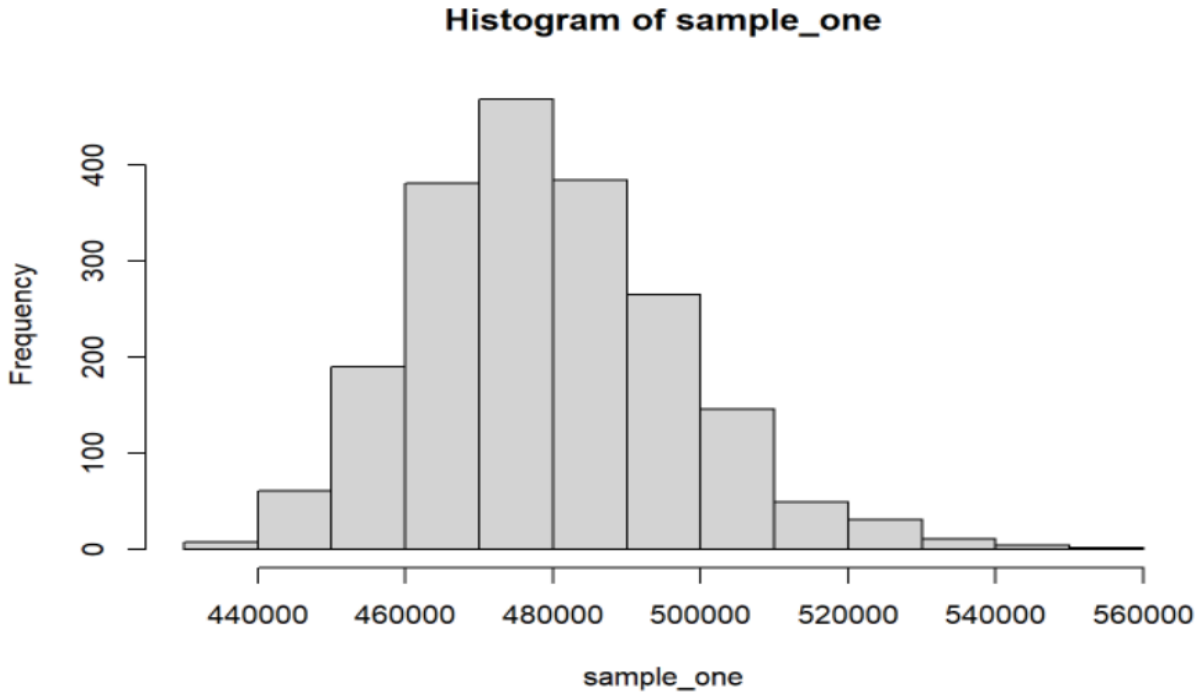

Figure 6. Histogram of the second sampling distribution