Project: Question 6

Gabriel Berardi

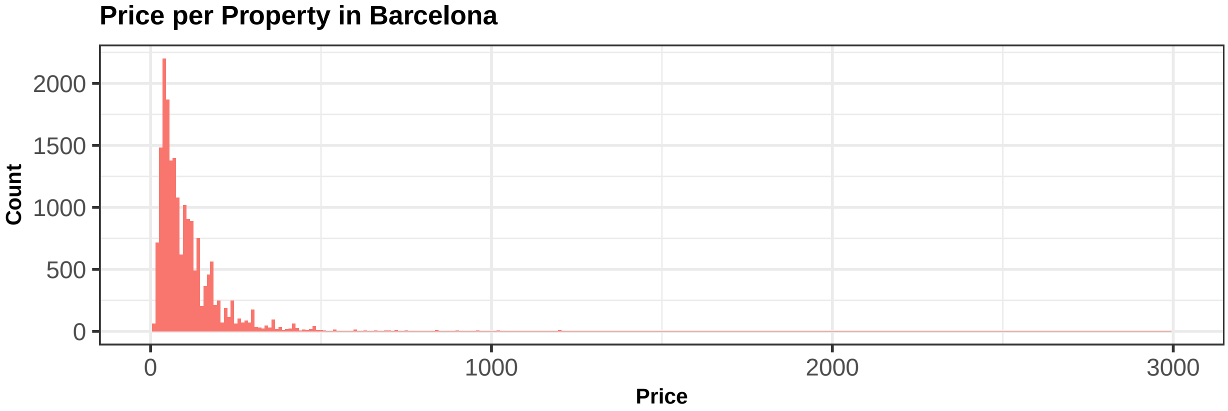
1. Summarise the Data in the Price Column

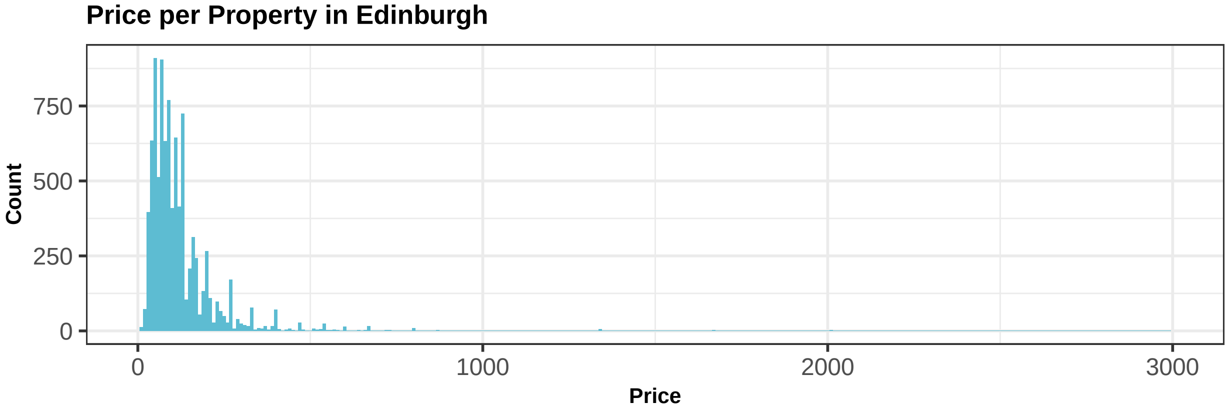
Let's take a closer look at the price column of the AirBnb listings in Barcelona and Edinburgh:

Table 1: Price per Property

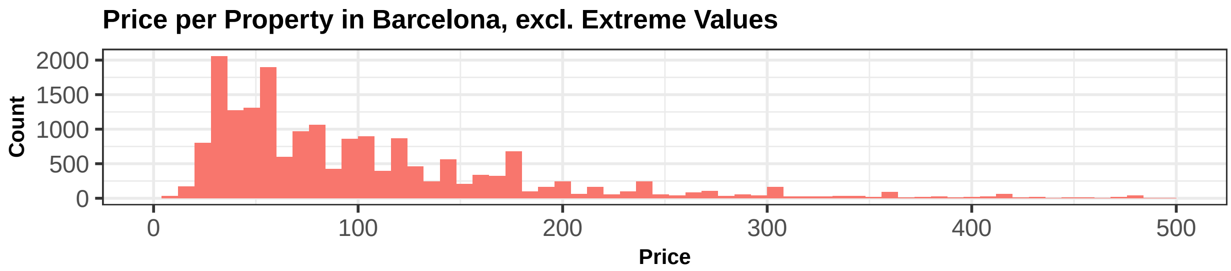
City	Minimum	LQ	Median	Mean	UQ	Maximum
Barcelona	10	48	78	110.2725	136	3606
Edinburgh	10	60	94	121.6147	138	2006

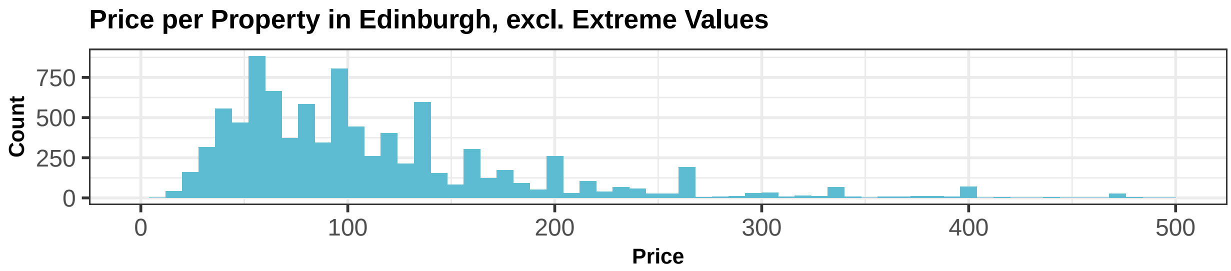
As we can see, the data for the AirBnb prices have a very large range and since the Mean lies to the right of the Median, we can tell that the data seems to be right-skewed. Let's investigate this with a histogram:



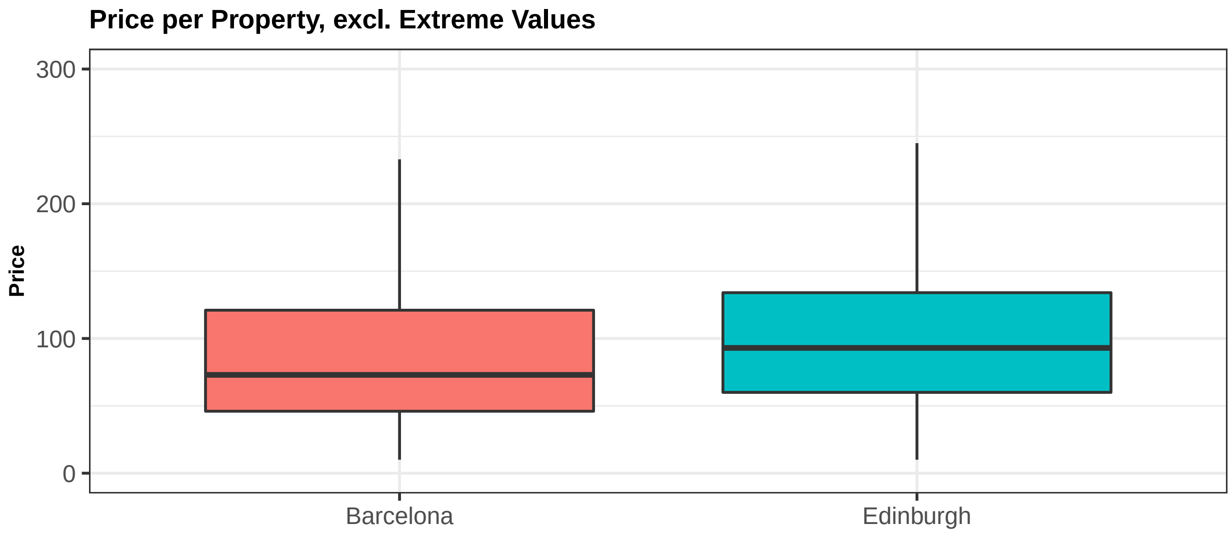


We can see that there are some very extreme values both in Barcelona and Edinburgh. As has been outlined before, these might be properties whose price data were captured for full months. Let's remove data that lie above the 99%-quantile and re-plot the histogram:





After we have removed the extreme values, we can see that the distribution for the price per property is quite similar in Barcelona and Edinburgh, but this can be investigated better using a boxplot:



Now we can see more clearly that the AirBnb prices in Edinburgh (Median: 94) appear to be higher than in Barcelona (Median: 78), when we ignore prices that lie above the 99%-quantile.

Let's now move on to investigate this a bit further with a one-sample confidence interval for the population mean price.

2. One-sample Confidence Interval for the Population Mean Price

First of all, let's think about the assumptions for one-sample confidence intervals and one-sample t-tests. The two main assumptions are:

1. our data x1 . . . , xn have arisen from a normal distribution;
2. our data points x1 , . . . , xn are independent of one another.

As we have noticed before, the data for price are highly right-skewed. However, since our sample for Barcelona and Edinburgh are both very large, we can still produce reasonable one-sample confidence intervals and perform useful one-sample t-tests. This is because we have a large sample for both cities and the central limit theorem tells us that the distribution of the prices, in repeated sampling, converges to a normal distribution, irrespective of the original distribution. **Nevertheless, the central limit theorem is still an approximation and therefore we will later also use a transformation on the price data to make it more normally distributed and use this to compare the results with and without transformation!**

With respect to the second assumption, we might need to consider whether the price of an AirBnb listing could be influenced by other listings, for example because homeowners compare and adjust their own listing's price to those nearby. However, for the remainder of this project, we will treat the data as though individual listings are not related to one another.

Let's first produce a one-sample confidence interval including all price data we have. I will do this for a confidence level of 95%, 99% and 99.9% and gather the results in a table.

```
##
##  One Sample t-test
##
## data:  barcelona$price
## t = 127.87, df = 18837, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  108.5821 111.9628
## sample estimates:
## mean of x
##  110.2725
```

```
##
##  One Sample t-test
##
## data:  barcelona$price
## t = 127.87, df = 18837, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  108.0509 112.4941
## sample estimates:
## mean of x
##  110.2725
```

```
##
##  One Sample t-test
##
## data:  barcelona$price
## t = 127.87, df = 18837, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99.9 percent confidence interval:
##  107.4343 113.1106
## sample estimates:
## mean of x
##  110.2725
```

```
##
##  One Sample t-test
##
## data:  edinburgh$price
## t = 106.31, df = 9401, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  119.3723 123.8570
## sample estimates:
## mean of x
##  121.6147
```

```
##
##  One Sample t-test
##
## data:  edinburgh$price
## t = 106.31, df = 9401, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  118.6675 124.5618
## sample estimates:
## mean of x
##  121.6147
```

```
##
##  One Sample t-test
##
## data:  edinburgh$price
## t = 106.31, df = 9401, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99.9 percent confidence interval:
##  117.8493 125.3800
## sample estimates:
## mean of x
##  121.6147
```

Table 2: Confidence Intervals for Population Mean Price

City	X0.95	X0.99	X0.999
Barcelona	(108.58, 111.96)	(108.05, 112.49)	(107.43, 113.11)
Edinburgh	(119.37 123.86)	(118.67, 124.56)	(117.85, 125.38)

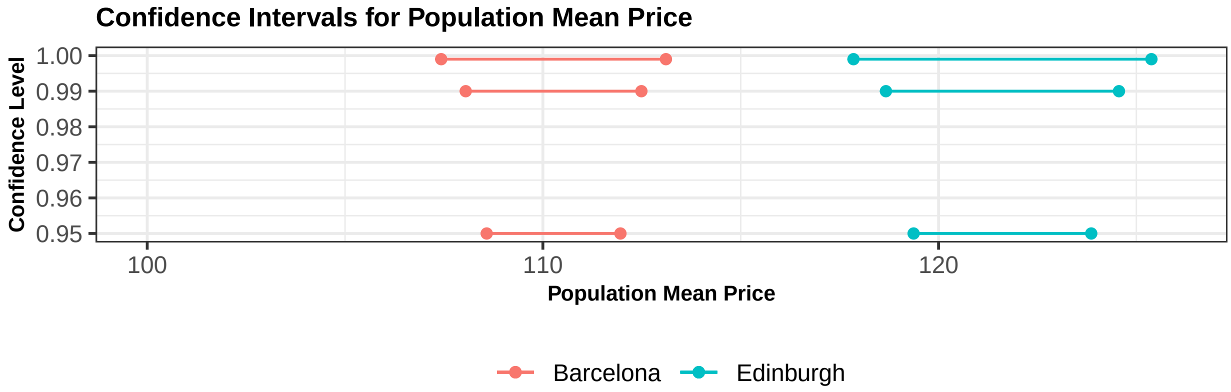
As we can see in Table 2, each of the 95 %, 99 % and the 99.9 % confidence interval for the population mean price show a clear gap between Barcelona and Edinburgh, with Edinburgh having higher prices than Barcelona.

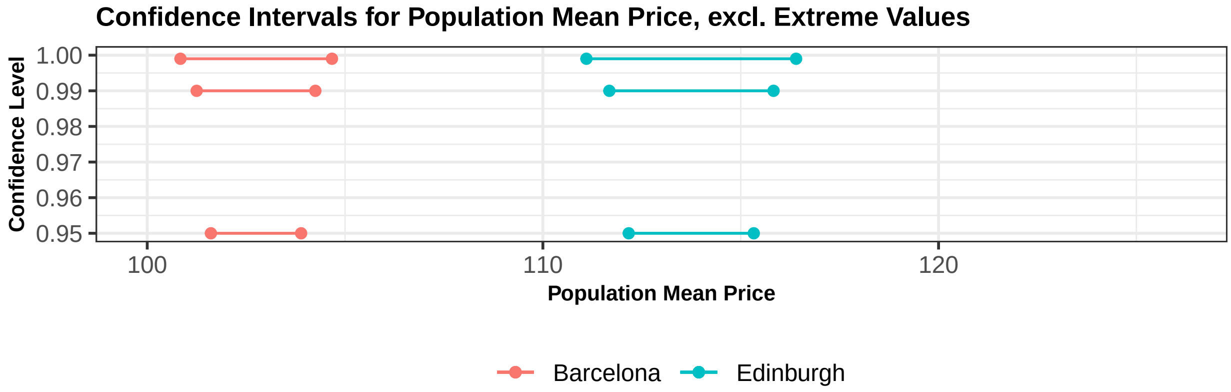
Now, we will attempt to account for the fact that some of the data for the AirBnb listings were recorded for monthly prices. As above, we will remove the listings whose price is higher than the 99% quantile, redo the one-sample tests and once again gather the data in a table. Here, I will omit the individual outputs and only show the final table:

Table 3: Confidence Intervals for Population Mean Price, excl. Extreme Values

City	X0.95	X0.99	X0.999
Barcelona	(101.61, 103.89)	(101.25, 104.25)	(100.84, 104.67)
Edinburgh	(112.17, 115.33)	(111.68, 115.83)	(111.10, 116.40)

After removing the extreme values for price, the distinction between Barcelona and Edinburgh is equally clear. To conclude this, let's plot these confidence intervals:



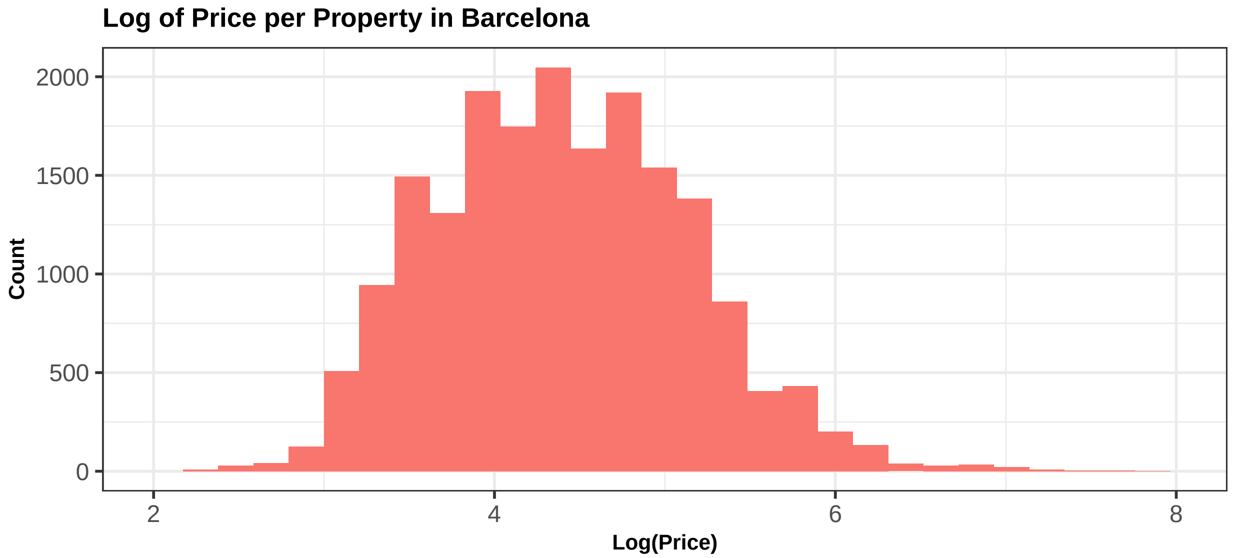


From these two plots above we can conclude two things:

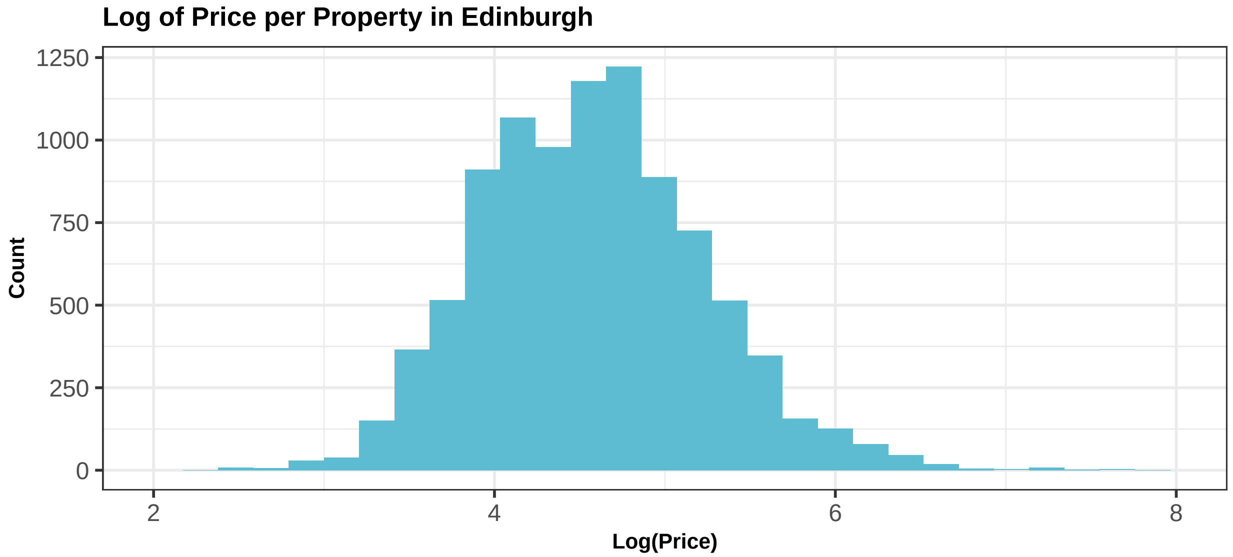
1. The population mean price seems to be lower in Barcelona across all confidence levels and both including and excluding extreme values.
2. When excluding extreme values, the confidence intervals for Barcelona and Edinburgh become narrower across all confidence levels.

As mentioned above, we will now apply a log-transformation on our price data and repeat the one-sample t-test.

Let's look at a histogram of the price data after log-transformation:



Log of Price per Property in Barcelona



Log of Price per Property in Edinburgh

As we can see, using a log-transformation on our data makes it much more normally distributed. Now, let's reproduce the one-sample t-test with a 95 % confidence level:

```
##
##  One Sample t-test
##
## data:  log(barcelona$price)
## t = 818.11, df = 18837, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.39291 4.41401
## sample estimates:
## mean of x
##   4.40346
```

```
##
##  One Sample t-test
##
## data:  log(edinburgh$price)
## t = 676.73, df = 9401, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.551792 4.578238
## sample estimates:
## mean of x
##  4.565015
```

Table 4: Confidence Intervals for Population Mean Log(Price)

City	X0.95
Barcelona	(4.39, 4.41)
Edinburgh	(4.55, 4.58)

Again, from Table 4 we can see that there is a clear distinction between the logarithm of the mean price in Barcelona and Edinburgh, as the two intervals do not overlap.

We can undo the log-transformation and obtain:

Table 5: Confidence Intervals for Population Geometric Price

City	X0.95
Barcelona	(80.64, 82.27)
Edinburgh	(94.63, 97.51)

Note that Table 5 now displays the 95 % confidence interval for the **geometric mean** of the prices in Barcelona and Edinburgh, which does make sense here, since the geometric mean is not overly influenced by the very large values in a skewed distribution.

This means, that we would also conclude that the geometric mean price for AirBnbs is higher in Edinburgh than in Barcelona.

3. Two-sample t-test for the Population Mean Prices

Now we will investigate the hypothesis that the population mean prices are different for Barcelona and Edinburgh. We now make the additional assumption that the distribution of the mean price has the same variance in Barcelona and Edinburgh. We can investigate this assumption with Levene's test for homogeneity of variance. In this case, our null hypothesis is that the population variance of price data is the same for AirBnb listings in Barcelona and Edinburgh:

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1    2.23 0.1354
##       28238
```

Since the p-value for Levene's test is well above 0.05, we would not reject the null hypothesis and conclude that we can assume that the population variances for price to be equal in Barcelona and Edinburgh.

We then have:

$H_0 : \mu_B = \mu_E$

$H_1 : \mu_B \neq \mu_E$

where μ is the population mean price for Barcelona and Edinburgh.

Unlike before, we will now only work with a confidence level of 95%.

```
##
##  Two Sample t-test
##
## data:  bande$price by bande$city
## t = -7.7476, df = 28238, p-value = 9.681e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.211623  -8.472727
## sample estimates:
## mean in group Barcelona mean in group Edinburgh
##                110.2725                121.6147
```

As we can see, the p-value is very small and well below 5 %, so there is clear evidence that we should reject our null hypothesis, i.e. that there is a statistically significant difference in the population mean price for AirBnb listings in Barcelona and Edinburgh at a significance level of 5 % (indeed, looking at the p-value, also at a much lower significance level). If we look at the confidence interval of (-14.21, -8.47), we can also say that the population mean price seems to be lower in Barcelona than in Edinburgh and the difference is likely to lie between 8.47 and 14.21.

Similarly as before, let's remove the extreme price values that lie above the 99 % quantile and repeat this two-sample test.

```
##
##  Two Sample t-test
##
## data:  bande_corrected$price by bande_corrected$city
## t = -9.9183, df = 27954, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.981316  -8.027221
## sample estimates:
## mean in group Barcelona mean in group Edinburgh
##                103.3463                113.3506
```

After removing the extreme values for price, we can see that the null hypothesis should still be rejected at a significance level of 5 %. In fact, the p-value is even lower than before. The confidence interval is narrower and the difference in population mean prices between Barcelona and Edinburgh is now estimated to lie somewhere between 8.03 and 11.98, with Barcelona having lower mean prices than Edinburgh.

Last but not least, let's again apply a log-transformation on the data and reproduce the two-sample t-test:

```
##
##  Two Sample t-test
##
## data:  log(bande$price) by bande$city
## t = -17.977, df = 28238, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1791687 -0.1439404
## sample estimates:
## mean in group Barcelona mean in group Edinburgh
##                4.403460                4.565015
```

As we can see, we still reject the null hypothesis and estimate that the difference in population mean log(prices) between Barcelona and Edinburgh lies somewhere between 0.14 and 0.18 with Barcelona having lower mean log(prices) than Edinburgh.