Chapter 7: Use Case Notebook for Instuctors

Ram Gopal, Dan Philps, and Tillman Weyde

Summer 2022

Contents

Use Case: Outlier Detection in Product Data

Testing whether a sample is likely to have come from a given population is a powerful use of probability, as discussed above. Using the same tools, we can also express whether a certain sample is an outlier in our observations: known as Outlier detection. Outlier detection is critical for manufacturing processes, detecting components that fall outside of the required tolerances; it can be important in identifying fraudulent transactions in a FinTech environment, along with many other uses in business.

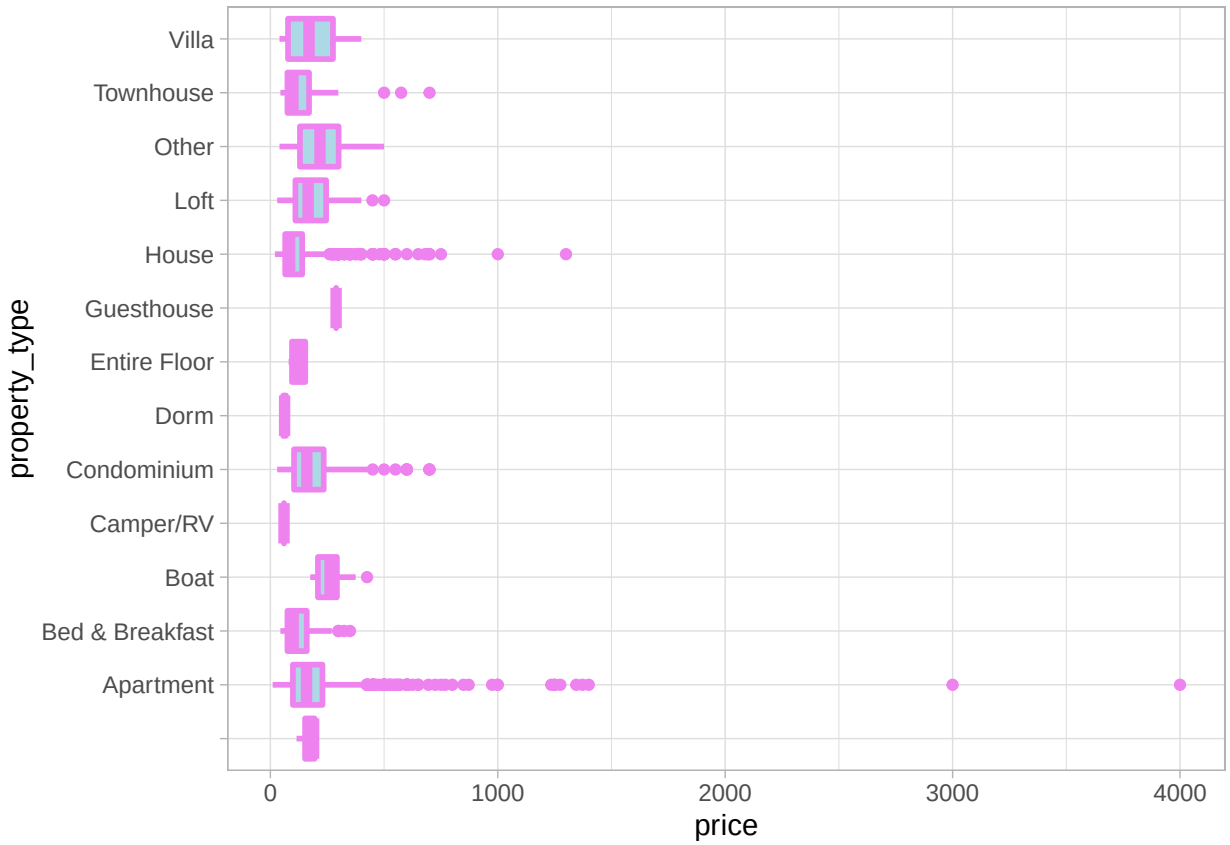
We can use the tools introduced above to identify outliers in a dataset, and once identified, we as business analysts, or managers, will need to determine whether the presence of outliers is acceptable (or indeed, beneficial).

```
df = read.csv("../../data/listings - wrangled.csv")
```

AirBnB Outliers in Boston

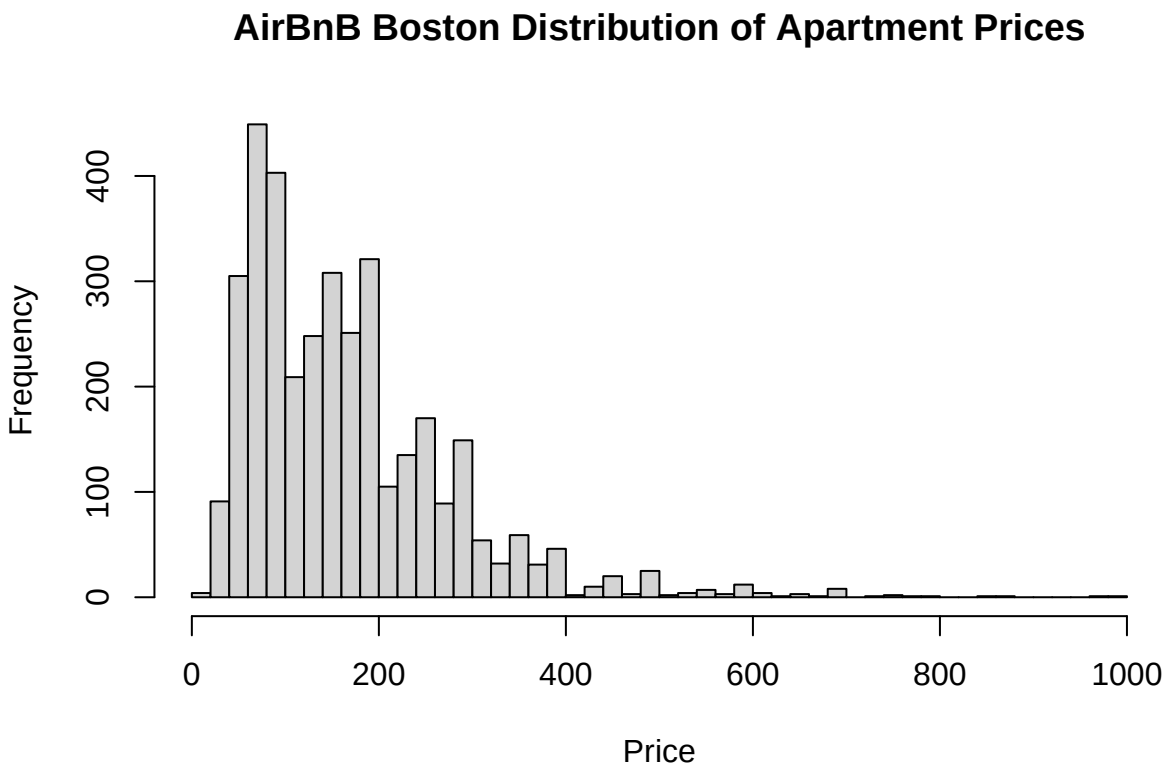
Taking the Boston Airbnb Open Data, which has listings, ratings and so on for different properties advertised and rented through AirBnB, we can examine the different properties on offer for outliers, that may or may not benefit AirBnB's offering. Below, you can see the distribution of price for different property types offered in Boston by AirBnB (note that we have to go through some data wrangling steps to clean the data ready for analysis):

```
library(ggplot2)
ggplot(df,aes(y = property_type ,x=price)) +
  geom_boxplot(col = "violet", fill = "lightblue", size = 1) +
  theme_light()
```



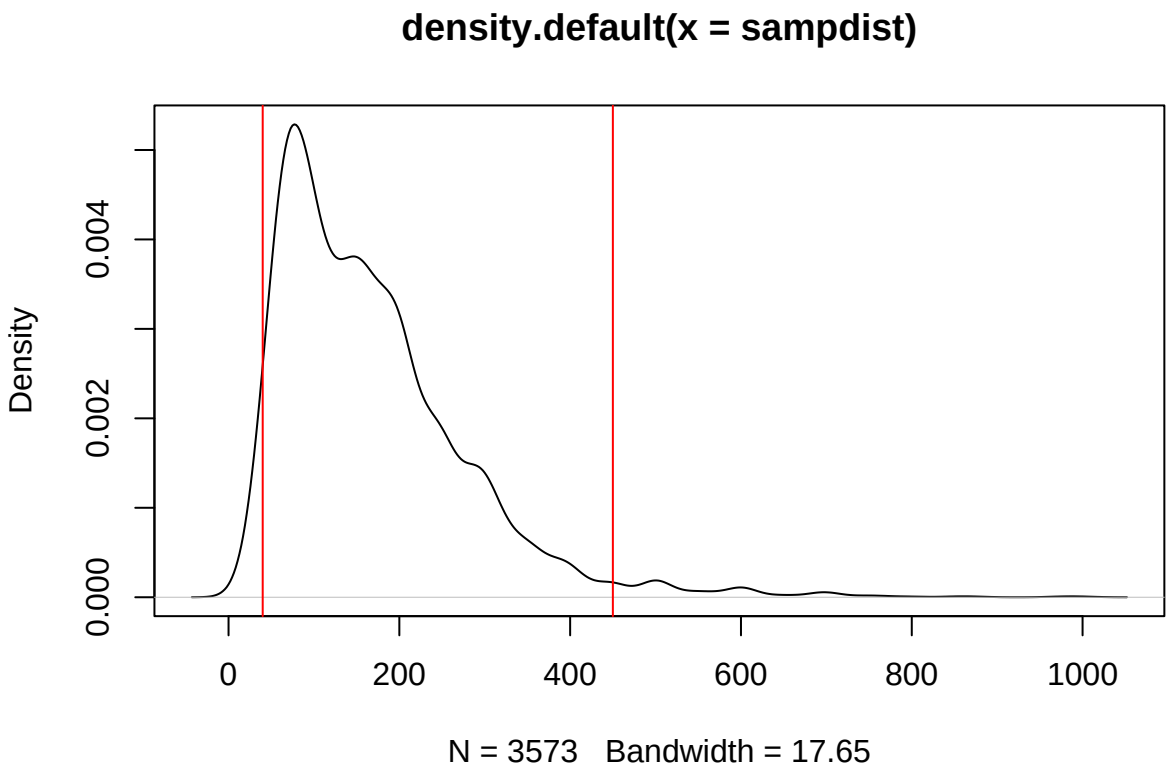
We can see from the chart that there are outliers in most categories, but particularly the Apartments category, where one apartment is offered for around $4000, whereas the median is $159. There are clearly a few very expensive apartments for rent on AirBnB in Boston. If we drill down into apartment prices, plotting a frequency distribution by binning apartment prices into 50 bins using a column chart, this helps visualize the distribution (note that we remove all outliers over $1000 in this case):

```
hist(df$price[df$price<1000],breaks=50,
     main = "AirBnB Boston Distribution of Apartment Prices",
     xlab = "Price")
```



Applying the confidence interval function we have learned above, conf_int, we can replot this data distribution with 95% confidence intervals:

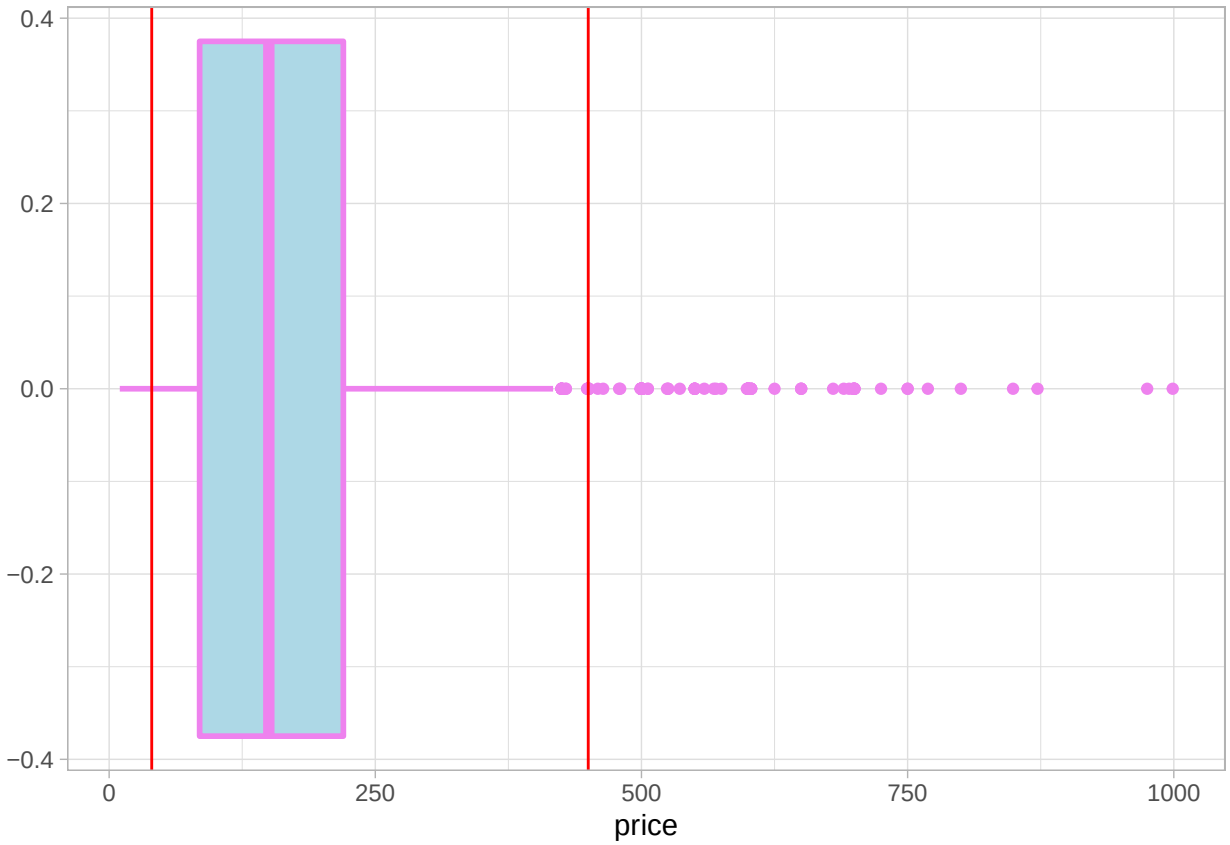
```
conf_int = function(sampdist,conlevel=0.95)
{

  left_v = (1 - conlevel)/2
  right_v = 1 - left_v
  q1 = quantile(sampdist,c(left_v,right_v))
  plot(density(sampdist))
  abline(v = q1, col = "red")
  output = paste0("[",q1[1],",",q1[2],"]")
  return(output)
}
conf_int(df$price[df$price<1000])
```




```
## [1] "[40,450]"
```

If we prefer to look at the distribution as a boxplot, we can see the extremes more clearly:

```
qvalues = quantile(df[df$price<1000,]$price,c(0.05/2,1-0.05/2))
ggplot(df[df$price<1000,],aes(x=price)) +
  geom_boxplot(col = "violet", fill = "lightblue", size = 1) +
  theme_light() +
   geom_vline(xintercept=qvalues,col="red")
```



It is certainly the case that the apartments renting for prices above the right-hand confidence interval or below the left-hand confidence interval, are significantly different offerings from the majority of the apartments on offer. They are outliers, and perhaps we can consider them different beasts – not from the same population as typical AirBnB rentals. It could be that these apartments rarely rent and represent an overhead for the platform. It could also be that they add prestige (and budget options) to the offering. But having identified these very different offerings statistically, we are now able to take whatever qualitative business actions are appropriate.