Group project - Data Management and Analysis: Coding R

Nguyen Khanh Chi & Tansum Tanha

2022-11-06

Focus of the study

For our group project on the hotel europe dataset, we decided to look at two cities: Berlin and Madrid. The reason we choose these two cities is that Berlin and Madrid are similar to each other in various aspects, for instance, they both are capital cities of two European countries, their size and population are comparatively closer than the others and both are popular tourist spots, we thought this could perhaps lead to a somewhat similar number of tourist accommodation types (On the dataset, Berlin has 4333 accommodation types and Madrid has 3814) and we were interested to see the similarities and dissimilarities in their price range, distance, star rating etc. We decided to keep all type of accommodations the data set has to make comments and comparisons for all of them.As we decide to focus majorly on price and how other factors (star rating, distance) may correlate with prices, for this study we would like to:

1. Conduct a general description of tourist accommodations in Berlin and Madrid, in terms of price, star rating, and distance.

2. Seek a general pattern between accommodation-star-ratings and their respective prices in Berlin and Madrid.

3. Spot any discernible pattern between distance from the city center and the accommodation prices**,** as well as their star ratings in Berlin and Madrid.

Fixing the data frame

We are working with accommodations situated in Berlin and Madrid in November 2017, we are leaving the weekends out to conduct the study with a total of 1039 observations.

Before we categorize accommodations based on star ratings, we take a look at how this variable varies among two cities. As per the table below, we classified accommodations into "Budget" (for accommodations rated up to 1 star); "Boutique" (rated from 1.5 to 2 stars); "Tourist" (rated from 2.5 to 3 stars); "Superior" (rated from 3.5 to 4 stars); and "Deluxe" (rated from 4.5 to 5 stars). accommodations are usually rated with half stars in the service industry, so we believe such classification is aligned with reality.

We also categorize the "distance" variable into "near" for accommodations that are within 2 miles from the city center; "moderate" for those that are between 2 to 4 miles from the city center; and "far" for those located more than 4 miles away from the city center. We excluded any NA values.

Comments on Descriptive Statistics

(Important: Due to technical reasons, we resorted to including the relevant graphs at the end of this document, under the "'Graph References" section. Please refer to those graphs, we apologise for any inconvenience caused and thank you for your understanding)

1. **Conduct a general description of tourist accommodations in Berlin and Madrid, in terms of price, star rating, and distance.**

a.) In terms of price

At first glance at the histograms (graph 1 & 2), we find that the distribution of hotel prices in Berlin and Madrid is skewed, with a long right tail, which means that there are several more values on the right end side of the ggplot that are further from the center value than the values on the left end. In general, the histogram shows that the accommodation price range in Berlin starts from 22 euros and only a few of the hotel prices are around 300 to 400 euros, and even fewer accommodations above 400 euros. The majority of accommodations in Berlin are priced from approximately from 80 to 110 euros per night. We can spot some extreme values: three places ranging from 270 to 330 euros per night, one charges 360 euros, one charges approximately 425 euros and one charges 480 euros. While Madrid's distribution of prices are mostly around 100 euros, and the price is spread out up to 300 euro per night with a few accommodations around 300 to 400 euro, and even fewer above 400 euro. In particular, the majority of tourist places in Madrid charge 60-80 euros per night, which is lower than Berlin. We can also spot some extreme values in Madrid accommodations: two places charge 300-310 euros per night, roughly 3-4 places charge almost 400 euros, one charges 490 euros and one costs nearly 600 euros, even higher than the maximum price in Berlin.

The boxplot (graph 3) compares the difference in price distribution among accommodations between Berlin and Madrid, supplemented by the descriptive table 9 and 10. The median price of accommodations offered in Berlin and Madrid is roughly the same at approximately 100 euros per night, which very much corresponds to their respective mean (100.32 and 110.12 euros), but places in Berlin have much smaller interquartile range of 40 euros compared to Madrid of 80 euros. Such difference in interquartile range is contributed by Madrid's higher upper quartile of 150 euros, while two cities share the same lower quartile of 70 euros. Berlin also has a longer lower whisker (min value) extending to 22 euros compared to Madrid's 30 euros per night, whereas Madrid has a longer upper whisker of 230 compared to Berlin's 195 euros per night. Table 9 and 10 also reveals that prices charged by accommodations in Madrid vary a lot more compared to Berlin, as prices in the former range from 574 to 30 euros as opposed to the latter, whose price per night vary from 22 to 470 euros.

Table 1: Berlin accommodations

	Mean	Max	Min	Range
price	100.32	470	22	448

Table 2: Madrid accommodations

	Mean	Max	Min	Range
price	110.12	574	30.00	544

b.) In terms of star-rating

Based on the descriptive table 1, of all 524 accommodations that Berlin has, nearly half of them are between 2.5 to 3 stars ("Tourist"), followed by those Superior, rated between 3.5 and 4 stars (roughly 34%), while only 13 accommodations are rated up to 1 star. In Madrid, accommodations rated 2.5 to 4 stars also make up the majority of their 515 places (more than 63% altogether), whereas Deluxe places (having stars from 4.5 to 5) share the least proportion (only approximately 5.6%). In general, tourists in both cities can most easily find either Tourist category accommodations (2.5~3 stars) or Superior (3.5~4 stars). On the other hand, tourists in Berlin and Madrid can face difficulties finding a good amount of Budget accommodation (up to 1 star) and Deluxe places (4.5~5 stars). We think this phenomenon is self-explanatory: most accommodations offer fall in the middle ground (2.5~4 stars), while tourists need to invest more time and effort to search in order to find Budget places. Deluxe accommodations cater for a very distinctive niche market, hence their limited availability. The mean star ratings of two cities are also very similar (3.32 and 3.27 for Berlin and Madrid respectively)

Class	Berlin	Madrid
Boutique (1.5 to 2 stars)	60	125
Budget (up to 1 star)	13	35
Deluxe (4.5 to 5)	51	29
Superior (3.5 to 4 stars)	178	182
Tourist (2.5 to 3 stars)	222	144
Mean star-rating	3.32	3.27

Table 3.Classification of tourist accommodations in Berlin and Madrid

c.) In terms of distance

According to the descriptive table below, we can say that most accommodations in Berlin and Madrid cluster around the city center, especially in Madrid (nearly 90% of accommodations in Madrid are within 2 miles from the city center, compared to 54.4% in Berlin). Accommodations further than 4 miles from the city center in both countries are hardly found.

Location	Berlin	Madrid
Far	18	4
Moderate	185	45
Near	242	412

Table 4.Number of accommodations

2. Seek a general pattern between accommodation-star-ratings and their respective prices in Berlin and Madrid.

We wanted to see if the star ratings and price indeed behave in the same way, so we created a descriptive table of accommodation prices factored by class. At first glance, such assumption holds its ground in both cities: on average, accommodations higher rated tend to charge high prices per night, with clear distinction between Deluxe/Superior accommodations and Tourist/Boutique/Budget places. In both cities, accommodations rated 3.5 and above cost over 100 euros per night, while those below this threshold cost less than 100. In Madrid, the price distinction is even more recognisable when Deluxe accommodations and penthouses' prices jump to 267.38 euros on average, while Superior places charge slightly more than 100 euros per night, and places under 3.5 stars prices remain steadily under 100 euros. Remarkably, except for Tourist places in Berlin, in both cities, there is a much larger range of price offers (320 to 301 euros difference in price for Deluxe and Superior accommodations respectively in Berlin, and 453 and 335 in Madrid) in accommodations above 3.5 stars compared to those below this threshold on average.

Table 5: Berlin accommodations

	Class	Mean	Max	Min	Range
price	Boutique (1.5-2 stars)	62.90	124	26	98
	Budget (~1 star)	70.69	102	22	80
	Deluxe (4.5-5 stars)	174.31	412	92	320
	Superior (3.5-4 stars)	111.84	359	58	301
	Tourist (2.5-3 stars)	85.86	470	27	443

Table 6: Madrid accommodations

	class	Mean	Max	Min	Range
price	Boutique (1.5-2 stars)	70.36	134	30	104
	Budget (~1 star)	61.66	101	34	67
	Deluxe (4.5-5 stars)	267.38	574	121	453
	Superior (3.5-4 stars)	137.66	389	54	335
	Tourist (2.5-3 stars)	90.12	179	36	143

3. Spot any discernible pattern between distance from the city center and the accommodation prices, as well as their star ratings in Berlin and Madrid

When we aim to comment on the relations between price and distance, we draw consistent conclusions between two cities. In detail, at a first glance, their tourist accommodation prices are averagely on par with each other with no concerning deviation. Obviously from the data, prices tend to become more expensive as the accommodations are nearer to the city center in both cities, with an exception in Madrid where the average hotel prices located near the city center (109.19) is actually lower than those within moderate distance ((114.02). The price range in Berlin is on the other hand higher than Madrid, however in both cities there is less difference in accommodations located far from the city center compared to those in near and moderate distance from the city center.

Table 7: Berlin accommodations

	location	Mean	Max	Min	Range
price	far	95.50	213	22	191
	moderate	99.32	470	26	444
	near	107.40	412	28	384

Table 8: Madrid accommodations

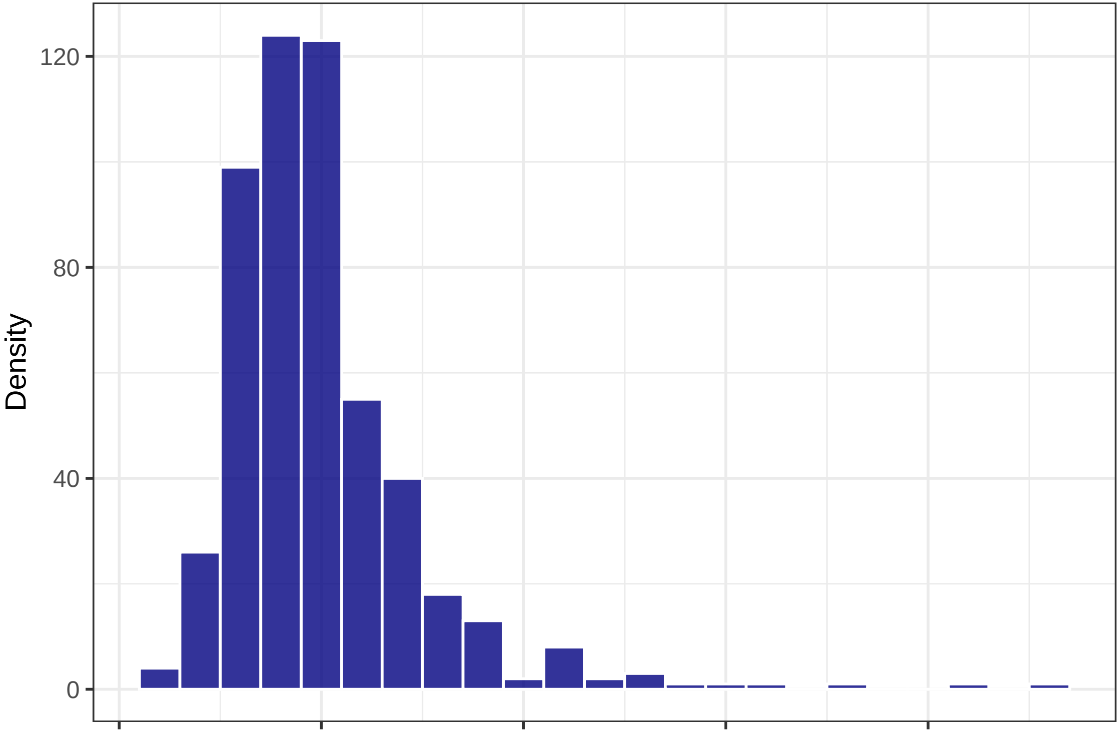
	location	Mean	Max	Min	Range
price	far	92	104	75	29
	moderate	114.02	389	39	350
	near	109.19	472	30	442

When we check the data summary on accommodation quality (based on star ratings) and distance, we find different results on average between two cities. In Berlin on average, Deluxe and Superior accommodations are located nearest to the city center while Budget and Tourist accommodations are the furthest, which understandably correlates with their price range. Remarkably, Superior accommodations seem to scatter all over the city - tourists can find Superior accommodations both near the center (0.6 miles away) and in the outskirts (13 miles away the furthest). On the contrary in Madrid, Boutique and Budget accommodations are located nearest to the city center on average (0.56 miles and 1.21 miles respectively) despite their cheapest prices, while Superior and Deluxe accommodations are the furthest from city center (averaging 2.01 and 1.33 miles respectively) as opposed to their costly price. We then defer that apart from distance, other factors must play a more important role in defining prices for tourist accommodations in Madrid. Notably, accommodations in Madrid are much more concentrated around the city center on average compared to Berlin.

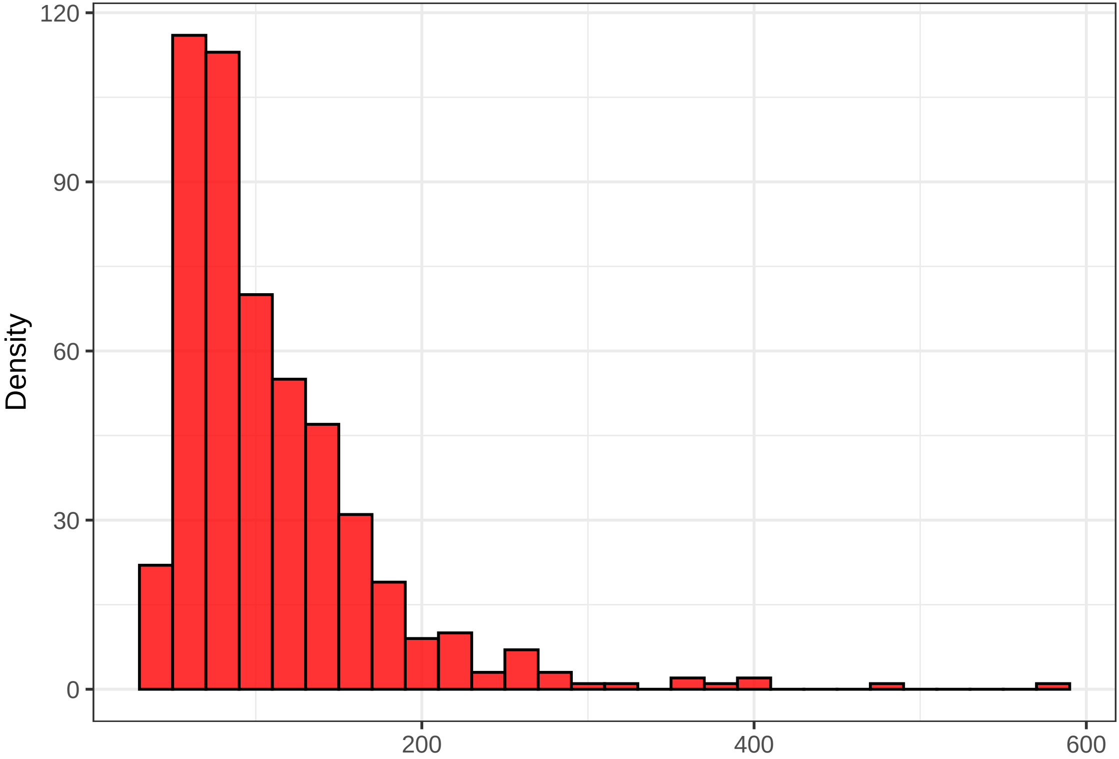
Complete codes
# Clear dataset

```
rm(list=ls())
#Importing libraries
library(tidyverse)
library(modelsummary)
library(ggthemes)
# Setting path
df <- read.csv(’https://osf.io/yzntm/download’))
# Two similar countries Data frame
Newdatafile <- df %>% filter(city_actual %in% c(‘Berlin’, ‘Madrid’))
filter(Newdatafile, year == 2017, month == 11 & weekend == 0)
TwoCities <- filter(Newdatafile, price <= 600, year == 2017, month == 11 & weekend == 0)
# Hotel Star rating
TwoCities <- TwoCities %>%
mutate(class = case_when(starrating == 0.0 | starrating == 1.0 ~ ‘Budget’,
starrating == 1.5 | starrating == 2.0 ~ ‘Boutique’,
starrating == 2.5 | starrating == 3.0 ~ ‘Tourist’,
starrating == 3.5 | starrating == 4.0 ~ ‘Superior’,
starrating == 4.5 | starrating == 5.0 ~ ‘Deluxe’))
xtabs(~ class + city_actual, data = TwoCities, addNA = TRUE)
# Distance
TwoCities <- separate(TwoCities, center1distance, ’ ’, into =
c(‘distance’, ‘miles’))
TwoCities <- select(TwoCities, -miles)
TwoCities$distance <- as.numeric(TwoCities$distance)
TwoCities <- filter(TwoCities, !is.na(TwoCities$distance))
TwoCities <- TwoCities %>%
mutate(location = case_when(distance == 0.0 | distance < 3.0 ~ ‘Near’,
distance == 3.0 | distance < 6.0 ~ ‘moderate’,
distance == 6.0 | distance > 9.0 ~ ‘far’))
xtabs(~ location + city_actual, data = TwoCities, addNA = TRUE)
#Separating cities
Berlin <- TwoCities %>% filter(city_actual %in% c(‘Berlin’))
Madrid <- TwoCities %>% filter(city_actual %in% c(‘Madrid’))
## Descriptive statistics
df2 <- Newdatafile %>%
group_by(starrating) %>%
summarize(mean = mean(price),
median = median(price),
min = min(price),
max = max(price))
print(df2)
# Ratings for diff cities
datasummary(starrating*city_actual ~ Mean, data=Newdatafile)
# Number of hotels in cities
count(Newdatafile, city_actual)
# Price-location datasummary for each countries
datasummary(price*location ~ Mean + Max + Min, data= Berlin)
datasummary(price*location ~ Mean + Max + Min, data= Madrid)
datasummary(price*class ~ Mean + Max + Min, data= Madrid)
datasummary(price*class ~ Mean + Max + Min, data= Berlin)
datasummary(distance*class ~ Mean + Max + Min, data= Berlin)
datasummary(distance*class ~ Mean + Max + Min, data= Madrid)
#ggplot for berlin
ggplot(filter(TwoCities, city_actual == ‘Berlin’), aes(x = price)) +
geom_histogram(alpha = 0.8, binwidth = 20, color=‘white’,
fill = ‘navyblue’) +
labs(x=‘Accomodation Prices in Berlin’,y=‘Density’)+
theme_bw()
#ggplot for madrid
ggplot(filter(TwoCities, city_actual == ‘Madrid’), aes(x = price)) +
geom_histogram(alpha = 0.8, binwidth = 20, color=‘black’,
fill = ‘red’) +
labs(x=‘Accomodation Prices in Madrid’,y=‘Density’)+
theme_bw()
#boxplot for both cities
f4 <- ggplot(TwoCities, aes(y = price, x = city_actual)) +
geom_boxplot(color = ‘blue’, size = 0.5, width = 2, alpha = 3) +
labs(x=‘Cities’,y=‘Price’) +
theme_bw()
f4
---
```

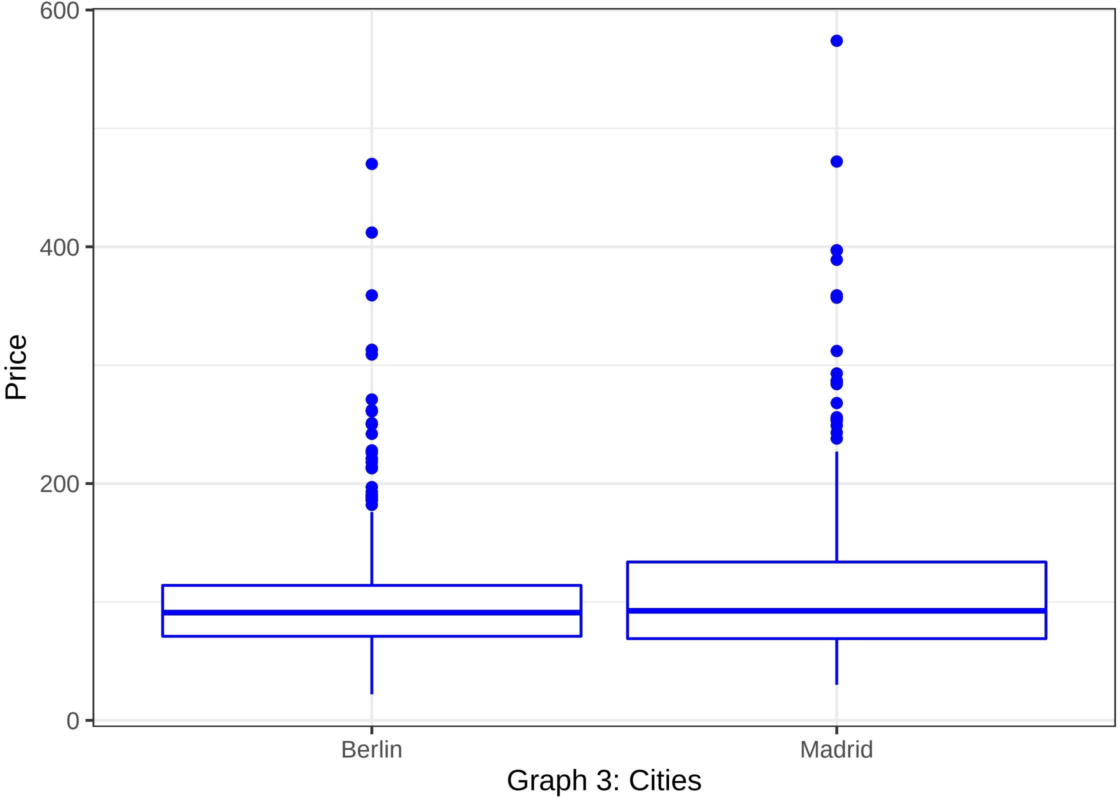
Graph References



Graph 1: Accomodation Prices in Berlin



Graph 2: Accomodation Prices in Madrid


600
400
200
0
Price
Berlin
Madrid
Graph 3: Cities