Group assignment

Alima Dzhanybaeva, Jana Hochel, Tasnim Murad

2022-05-11

1.1 Overview of the original data.

The original hotels-europe data includes information on price and features of hotels in 46 European cities and for 2017-2018. The data was downloaded from https://osf.io/yzntm/ download.

Table 1: Description of the main variables in the original dataset

variables	info	type
hotel_id	Hotel ID	numeric
accommodation_type	Type of accomodation	string
addresscountryname	Country	string
weekend	Flag, if day is a weekend	binary
holiday	Flag, if day is a public holiday	binary
center1distance	Distance from main city center	string
starrating	Number of stars	numeric
guestreviewrating	User rating average	string
price	Price in EUR	numeric
price_night	Number of nights	string
year	Year (YYYY)	numeric
month	Month (MM)	numeric

1.2 Dataset for 7 Central European countries

• **For our analysis we used 7 Central European countries: Czech Republic, Germany, Italy, Hungary, Austria, Poland, Slovakia.**

```
df2 <- df %>% filter(addresscountryname %in%
                       c('Czech Republic', 'Germany', 'Italy',
                         'Hungary', 'Austria', 'Poland', 'Slovakia'))
```

• **The folowing manipulations were carried out on the original data to change the existing variables and to add new ones:**

- The type of the variables *center1distance* and *guestreviewsrating* were changed from 'string' to 'numeric', and from 'string' to 'factor' for *addresscountryname*. The new variable names are *distance*, *actualrating* and *country* respectively.

- The original *accommodationtype* variable was transformed into new *acctype_f* factor variable.

```
df2 <- separate(df2, accommodationtype, '@', into =
                  c('word', 'acctype'))
df2 <- select(df2, -word)
df2 <- mutate(df2, acctype_f = factor(acctype))
```

- *trueprice* variable was generated by dividing the original *price* variable by *price_night*

```
df2 <- separate(df2, price_night, ' ', into =
                      c('pr','word', 'nights', 'night'))
df2 <- select(df2, -pr)
df2 <- select(df2, -night)
df2 <- select(df2, -word)
df2$nights <- as.numeric(df2$nights)
df2$trueprice <- df2$price/df2$nights
```

- *season* variable was created based on the *month* variable

```
df2 <- df2 %>%
  mutate(season = case_when(month == 12 | month == 1 |month == 2 ~ 'winter',
                            month == 3 | month == 4 |month == 5 ~ 'spring',
                            month == 6 | month == 7 |month == 8 ~ 'summer',
                            month == 9 | month == 10 |month == 11 ~ 'autumn'))
df2$season <- as.factor(df2$season)
```

- With the help of the existing *starrating* variable new *class* variable was generated

```
df2 <- df2 %>%
  mutate(class = case_when(starrating == 1.0 | starrating == 1.5 ~ 'tourist',
                           starrating == 2.0 | starrating == 2.5 ~ 'standard',
                           starrating == 3.0 | starrating == 3.5 ~ 'comfort',
                           starrating == 4.0 | starrating == 4.5 ~ 'first class',
                           starrating == 5.0 ~ 'luxury'))
df2$class <- as.factor(df2$class)
```

- **Quality of the dataset - missing values overview and frequency distribution of numeric values**

```
df2 %>%
  skimr::skim_without_charts()
```

Table 2: Data summary

Name	Piped data
Number of rows	54947
Number of columns	28
Column type frequency:	

Table 2: Data summary

character	8
factor	4
numeric	16
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
city_actual	0	1	3	33	0	278	0
center1label	0	1	11	11	0	1	0
center2distance	0	1	8	9	0	149	0
center2label	0	1	9	25	0	14	0
neighbourhood	0	1	3	63	0	427	0
s_city	0	1	4	10	0	13	0
acctype	0	1	0	19	31	16	0
offer_cat	0	1	11	13	0	5	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
country	0	1.00	FALSE	7	Ita: 27208, Ger: 9160, Pol: 5780, Aus: 5350
acctype_f	0	1.00	FALSE	16	Hot: 29880, Apa: 8878, Gue: 7311, Bed: 4814
season	0	1.00	FALSE	4	win: 22050, spr: 16298, aut: 10908, sum: 5691
class	15889	0.71	FALSE	5	com: 17695, fir: 14154, sta: 4025, lux: 2307

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
rating_reviewcount	4834	0.91	138.41	212.06	1	20.0	67.0	168.0	3234
distance	0	1.00	2.89	5.61	0	0.7	1.2	2.4	57
price	0	1.00	180.62	242.41	12	78.0	112.0	184.0	14859
nights	0	1.00	1.32	0.93	1	1.0	1.0	1.0	4
starrating	0	1.00	2.42	1.69	0	0.0	3.0	4.0	5
rating2_ta	6988	0.87	4.00	0.62	1	3.5	4.0	4.5	5
rating2_ta_reviewcount	6988	0.87	438.66	661.84	0	57.0	183.0	549.0	7717
actualrating	4834	0.91	3.98	0.58	1	3.7	4.0	4.4	5
scarce_room	0	1.00	0.64	0.48	0	0.0	1.0	1.0	1
hotel_id	0	1.00	13082.36	6385.14	1745	9933.0	14461.0	18293.0	22842
offer	0	1.00	0.55	0.50	0	0.0	1.0	1.0	1
year	0	1.00	2017.59	0.49	2017	2017.0	2018.0	2018.0	2018
month	0	1.00	6.86	4.12	1	3.0	6.0	11.0	12
weekend	0	1.00	0.66	0.47	0	0.0	1.0	1.0	1

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
holiday	0	1.00	0.21	0.41	0	0.0	0.0	0.0	1
trueprice	0	1.00	135.22	129.78	12	77.0	104.0	152.0	7674

- **Some value are not blank but include zeros and inf. The next table checks these entries as well as unique values.**

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
city_actual	0	0.00	0	0.00	0	0	character	278
rating_reviewcount	0	0.00	4834	8.80	0	0	integer	697
distance	236	0.43	0	0.00	0	0	numeric	143
center1label	0	0.00	0	0.00	0	0	character	1
center2distance	0	0.00	0	0.00	0	0	character	149
center2label	0	0.00	0	0.00	0	0	character	14
neighbourhood	0	0.00	0	0.00	0	0	character	427
price	0	0.00	0	0.00	0	0	integer	1317
nights	0	0.00	0	0.00	0	0	numeric	2
s_city	0	0.00	0	0.00	0	0	character	13
starrating	15889	28.92	0	0.00	0	0	numeric	10
rating2_ta	0	0.00	6988	12.72	0	0	numeric	9
rating2_ta_reviewcount	9	0.02	6988	12.72	0	0	integer	1467
acctype	0	0.00	0	0.00	0	0	character	16
actualrating	0	0.00	4834	8.80	0	0	numeric	27
scarce_room	19968	36.34	0	0.00	0	0	integer	2
hotel_id	0	0.00	0	0.00	0	0	integer	10396
offer	24910	45.33	0	0.00	0	0	integer	2
offer_cat	0	0.00	0	0.00	0	0	character	5
year	0	0.00	0	0.00	0	0	integer	2
month	0	0.00	0	0.00	0	0	integer	8
weekend	18754	34.13	0	0.00	0	0	integer	2
holiday	43219	78.66	0	0.00	0	0	integer	2
country	0	0.00	0	0.00	0	0	factor	7
acctype_f	0	0.00	0	0.00	0	0	factor	16
trueprice	0	0.00	0	0.00	0	0	numeric	1598
season	0	0.00	0	0.00	0	0	factor	4
class	0	0.00	15889	28.92	0	0	factor	5

- **Summary for the main variables in the transformed dataset**

Numeric variables

	Mean	Min	Max	SD
distance	2.89	0.00	57.00	5.61
starrating	2.42	0.00	5.00	1.69
actualrating	3.98	1.00	5.00	0.58
trueprice	135.22	12.00	7674.00	129.78

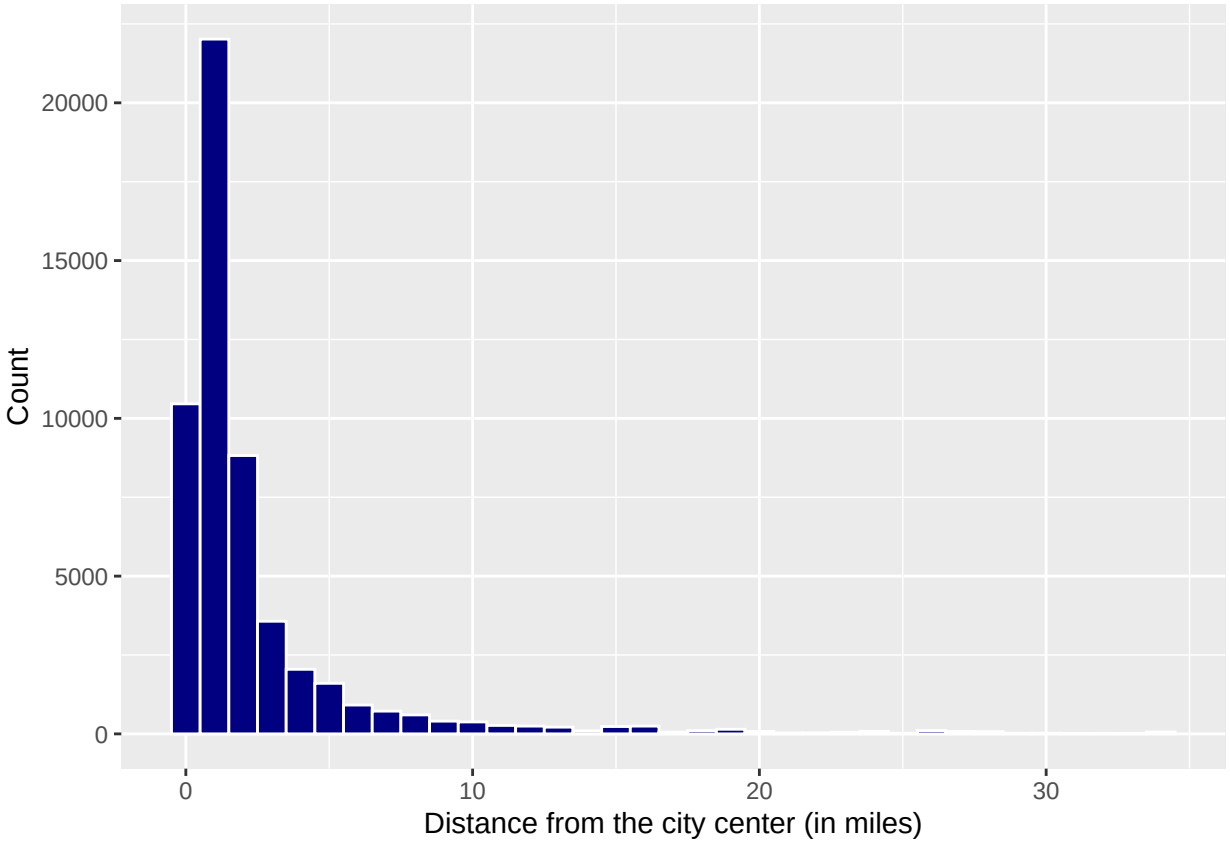
There are prominent outliers for *distance* and *true price* variables, therefore, the dataset was transformed, setting the maximum values at the 99% percentile (35 miles and 600 EUR respectively).

	mean	Min	Max	SD
distance	2.48	0.00	34.00	3.93
trueprice	126.49	12.00	599.00	82.03

- **Distance**

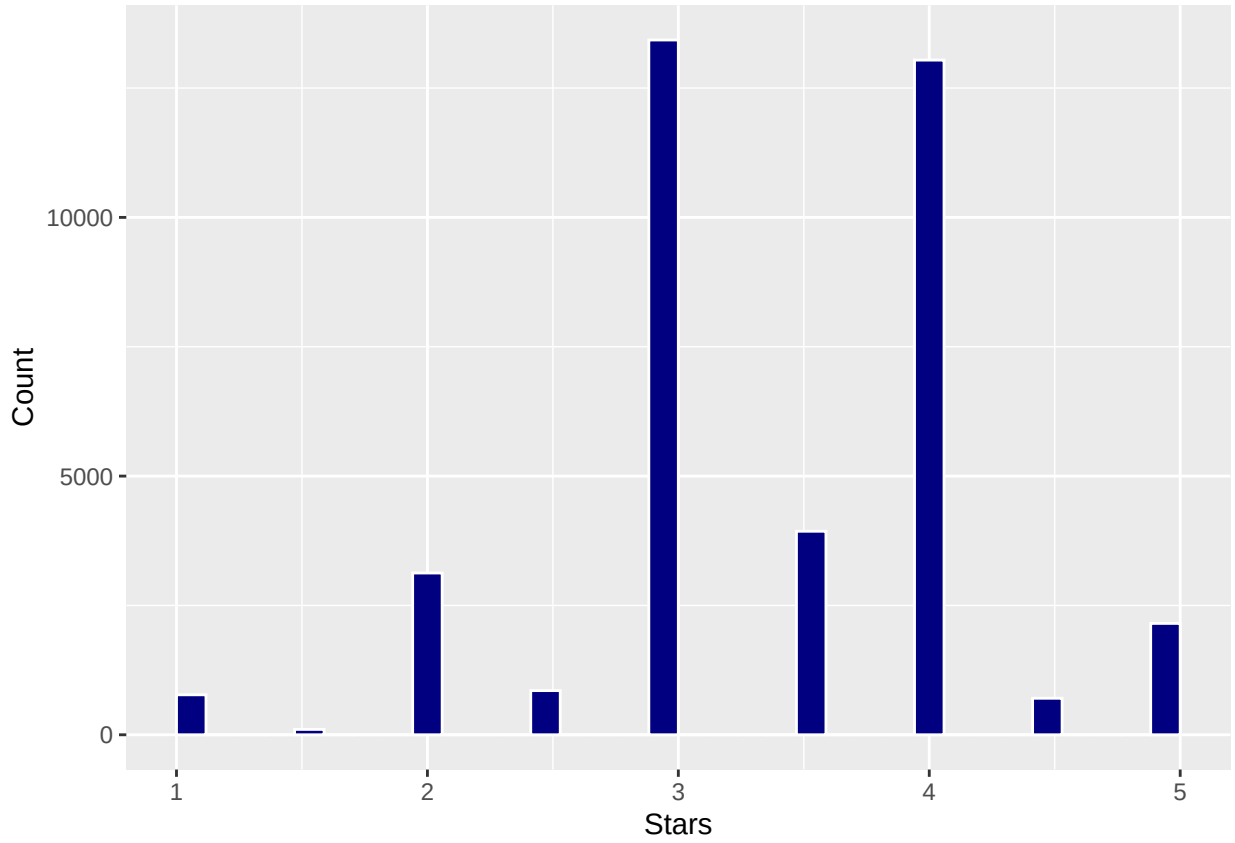
75% of the hotels are located between 0 and 2.4 miles from the main city center.

The histogram below shows the distribution of the distances from the city center.



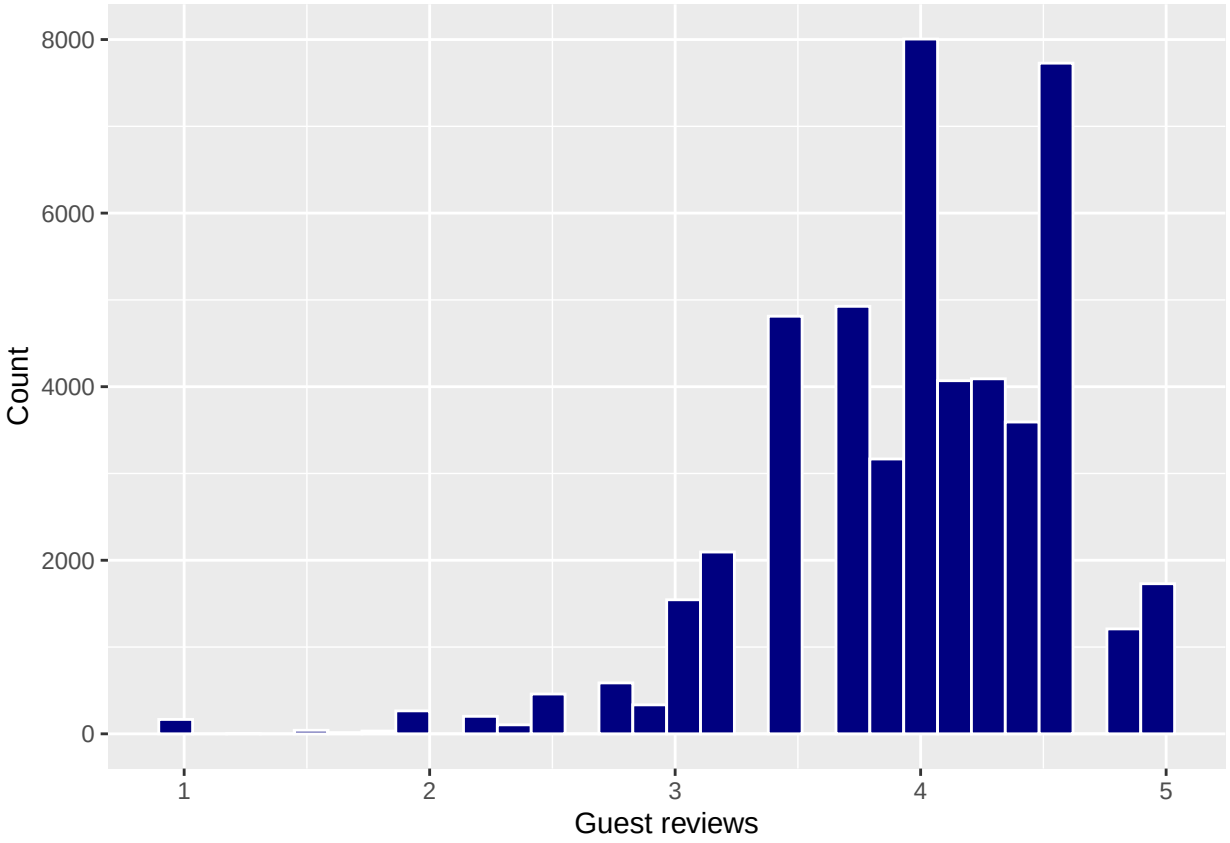
- **Starrating**

Most of hotels from the dataset have 3 and 4 stars.



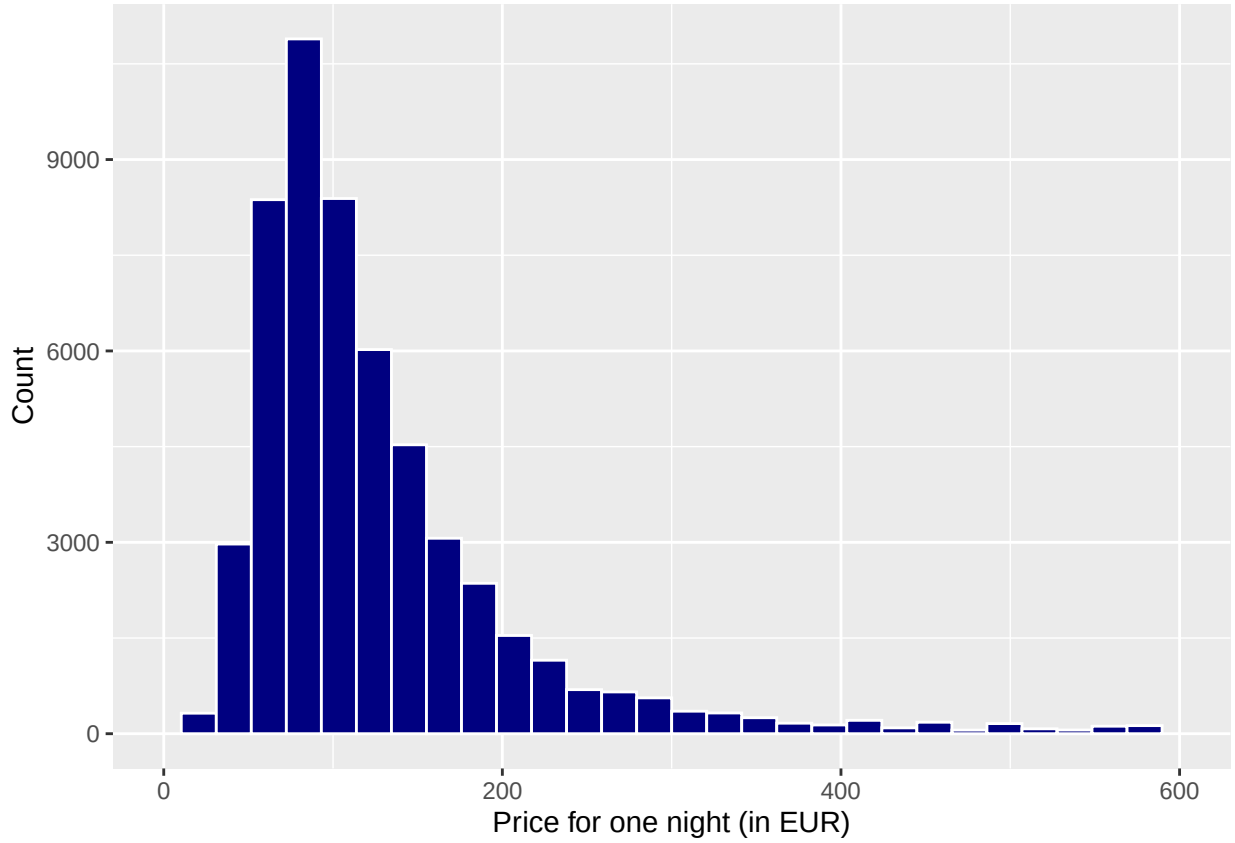
- **Actualrating**

As we can see from the histogram below, 80% of the guest reviews are between 3.4 and 4.1.



- **Trueprice**

The distribution of the prices is skewed to the right, therefore, we can conclude that the mean price is greater then the mode.



- **Year and month**

22270 records from the dataset are for the year 2017 and 31570 - for 2018.

year	n
2017	22270
2018	31570

month	n
1	4114
2	6037
3	5221
4	5226
5	5436
6	5536
11	10724
12	11546

Winter and spring are fully presented in the dataset, whereas there are no records for July, August, September, and October.

Factor variables

- **Accommodation type**

There are 16 unique accommodation types in the dataset. The most numerous types are "Hotel", "Apartment", "Guest House", and "Bed and Breakfast". Together they make up 93% of all records.

acctype_f	n
Hotel	29187
Apartment	8701
Guest House	7214
Bed and breakfast	4798
Hostel	1608
Apart-hotel	798
Pension	559
Inn	549
Vacation home Condo	311
	26
Caravan Park	26
Motel	25
Country House	18
House boat	17
Cottage	3

- **Country**

The countries with the most records are Italy and Germany. Together they make up 66% of all records.

country	n
Italy	26854
Germany	9123
Poland	5713
Austria	4775
Czech Republic	3834
Hungary	2697
Slovakia	844

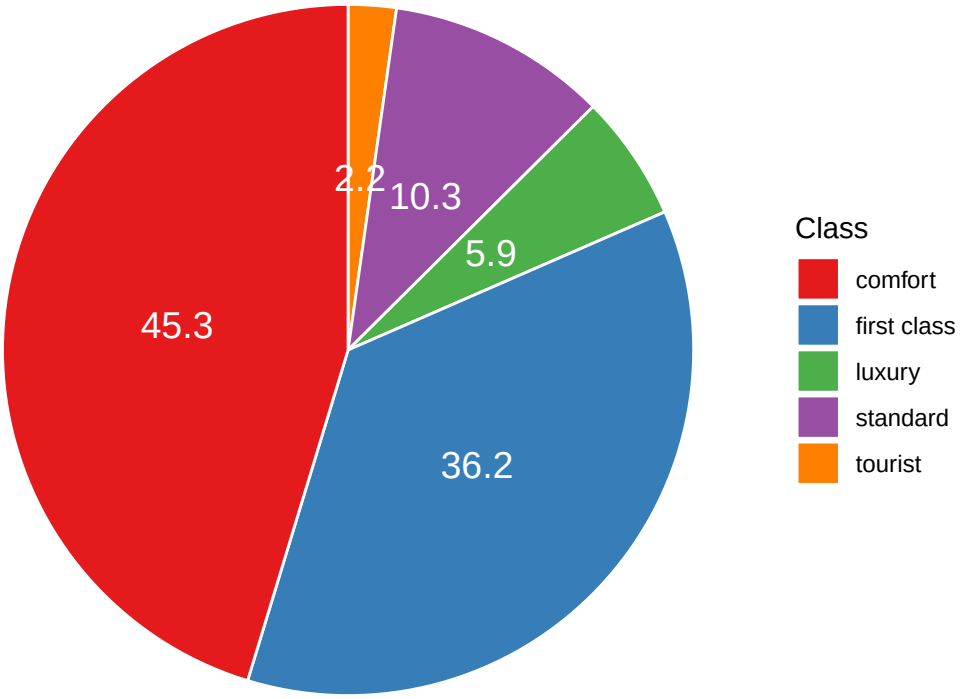
- **Season and class**

As we observed earlier, the dataset did not include two months for summer and two months for autumn, consequently, spring and winter have the most observations.

season	n
winter	21697
spring	15883
autumn	10724
summer	5536

The most obsevations belong to 'comfort' class, which corresponds to 3 and 3.5 stars, and to 'first class', which corresponds to 4 and 4.5 stars. There are also 15889 missing values. They appeared, because the hotels with 0 stars were not used for the 'class' variable, as the mean price for them is higher than for places with 1, 1.5, 2, and 2.5 stars, which does not make sense. Therefore, these observations were omitted.

class	n
comfort	17363
first class	13746
luxury	2153
standard	3981
tourist	877
NA	15720



Binary variables

- **Weekend and holiday**

35362 records were made on weekends (66%), and 18478 - on weekdays (34%)

weekend	n
0	18478
1	35362

Only 21% (11546) of all observations were made on holidays.

holiday	n
0	42294
1	11546

1.3 Analysis

For this project, we wanted to find the best deal among different types of accommodation in Central European countries.

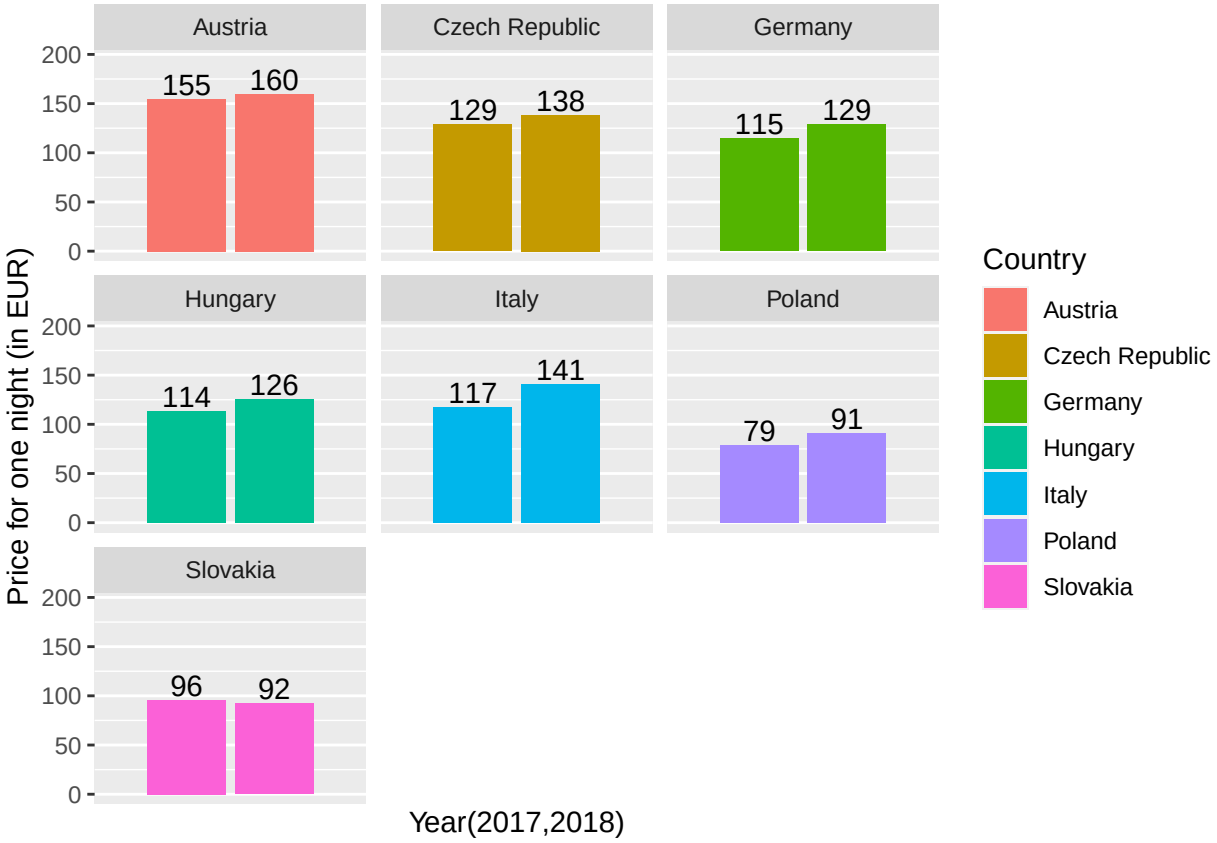
Consequently, we formulated the following question:

- *How does the average price for accommodation change for different values of the following variables: year, country, distance, season, class, weekend?*

Findings:

• As we can see from the graph below, the price increased for 6 countires, Slovakia being the only case in which the price dropped.

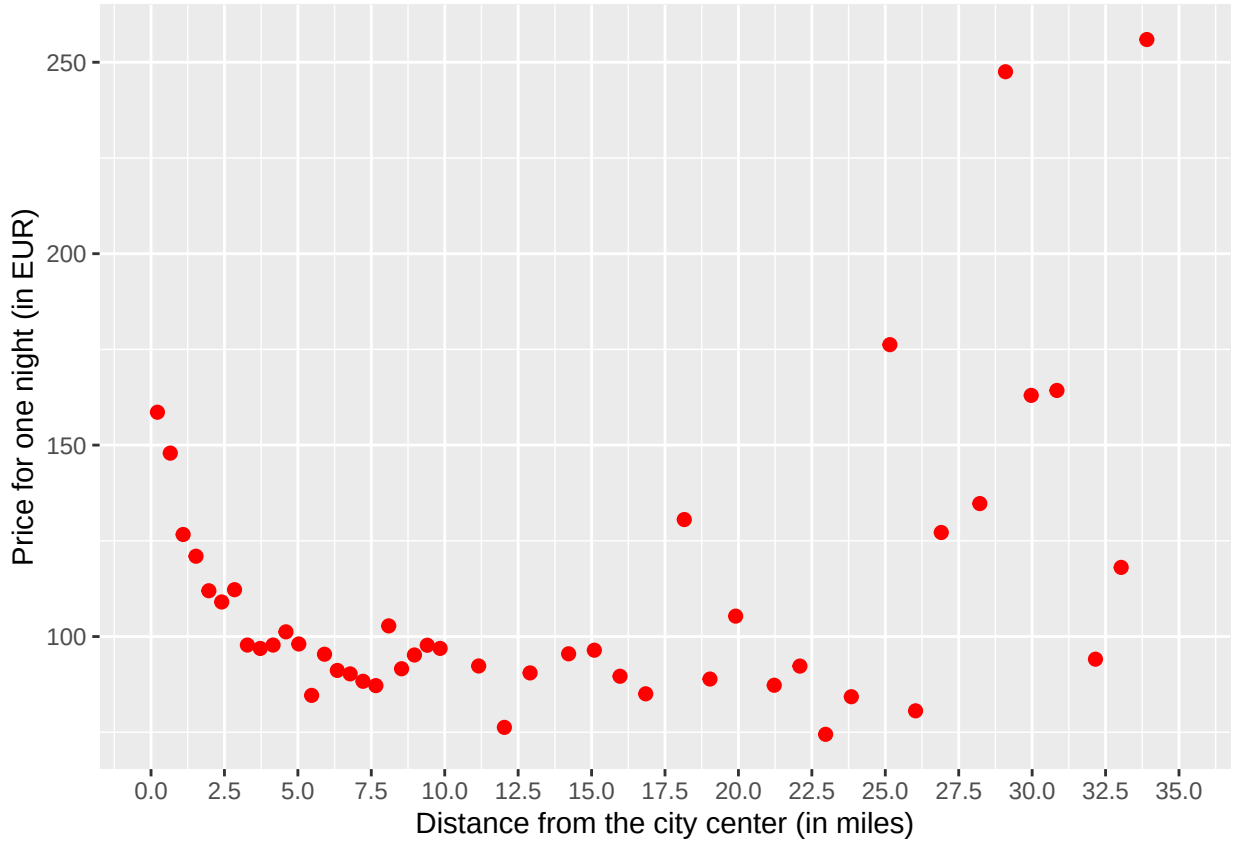
The biggest jump in price was in Italy. In this country, the price in 2018 increased by 24 euros compared to 2017.



• The lowest price is 12 EUR per night (Czech Republic), whereas the lowest average cost is 86 EUR (Poland). On average, Austrian accommodation is the most expensive with a mean price of 158 EUR.

	country	Mean	Max	Min
trueprice	Austria	157.74	598.00	27.00
	Czech Republic	134.28	588.00	12.00
	Germany	122.61	580.00	22.00
	Hungary	120.79	594.00	19.00
	Italy	131.34	599.00	16.00
	Poland	86.11	546.00	16.00
	Slovakia	93.59	573.00	20.00

• **The eminent negative relationship between price and distance from the city center can be tracked only until 7.5 miles. After that point, the dependence of the accommodation cost on the *distance* variable becomes vaguer, as the points are much more dispersed and do not follow a specific pattern. Therefore, we can conclude that on average, the cheapest accommodation is 7.5 miles far from the city center.**



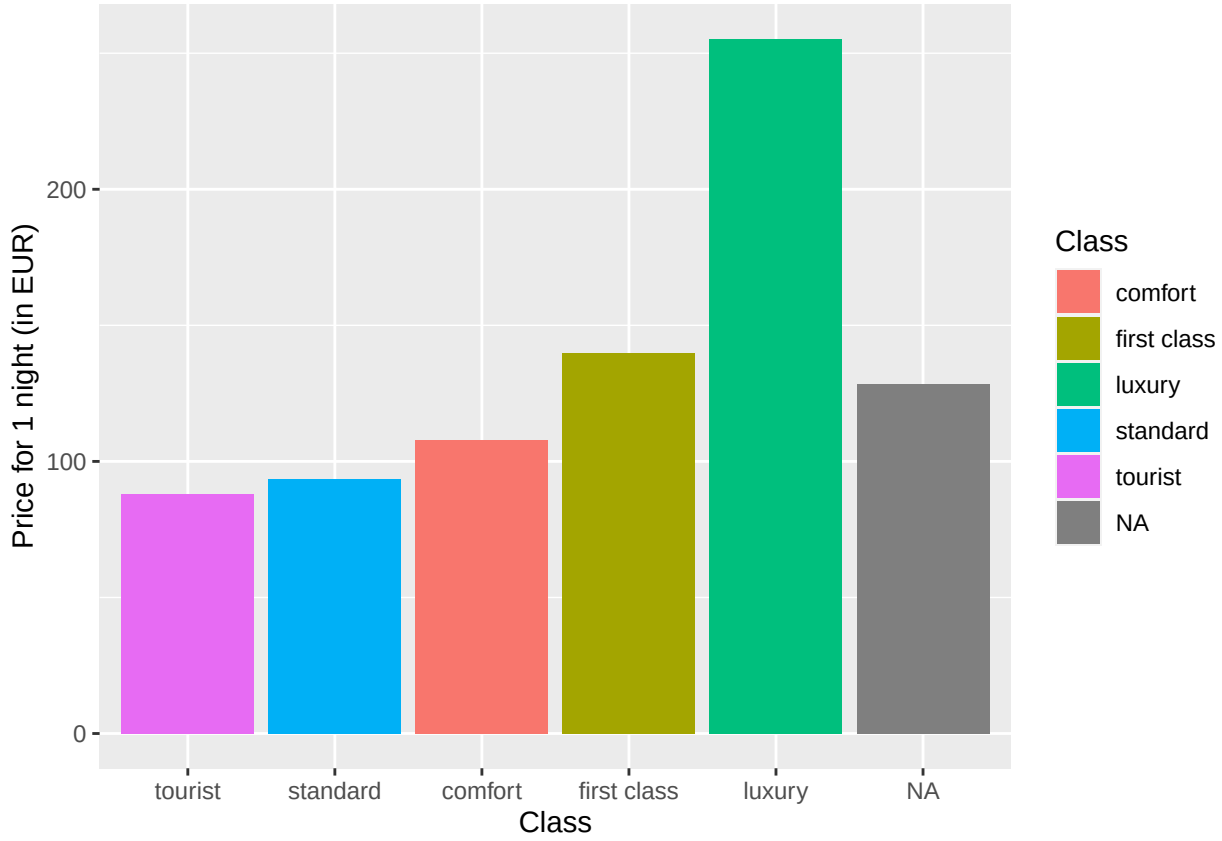
• **The prices for accommodation in the chosen 7 Central European countries are on average the lowest during autumn and are highest during spring. The difference between the lowest and the highest value is 35 euros.**

	season	Mean
trueprice	autumn	109.04
	spring	144.11
	summer	141.29
	winter	118.44

	class	Mean
trueprice	comfort	107.69
	first class	139.67
	luxury	255.33
	standard	93.41
	tourist	88.15

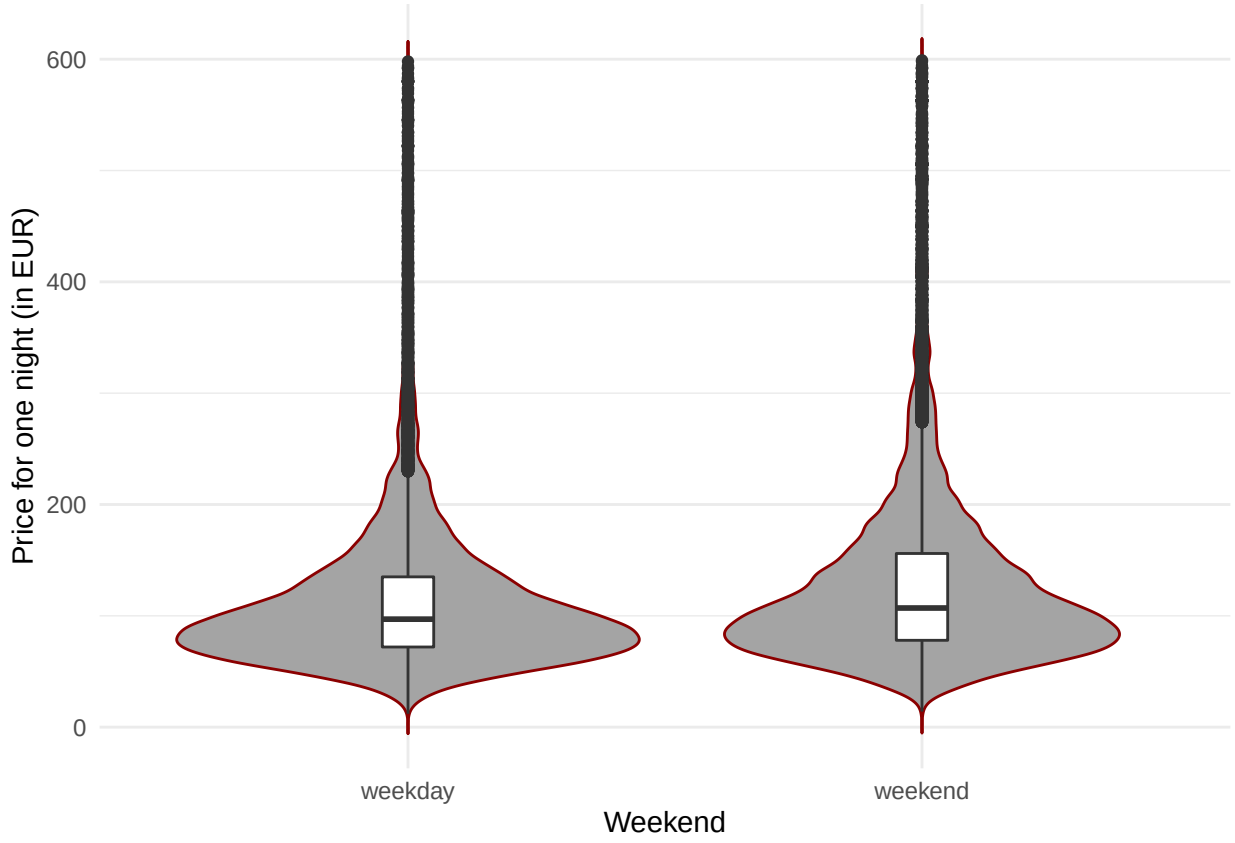
- **As expected, the cheapest class is "tourist" (1 and 1.5 stars) with a mean price of 88 euros. However, the average cost for "standard" (2 and 2.5 stars) falls not that far apart from the best deal with the difference of only 5 euros. Conversely, the "luxury" (5 stars) class is on average 140 euros more expensive than "first class" (4 and 4.5 stars).**

Consequently, the difference between mean prices becomes bigger with the number of stars.



	weekend	Mean
trueprice	weekday	115.65
	weekend	132.16

• The average price for accommodation is 16 euros higher during weekends. Moreover, as we can see from the graph below: (1) the boxplot for weekends is comparatively taller than the one for weekdays, indicating that prices are more dispersed during the end of the week; (2) the upper quartile for weekends is farther from the median than for weekdays, meaning that the 75% percentile is higher.



Summary

For the chosen 7 Central European countries (with the maximum value for the *distance* variable set to 35 miles and to 600 euros for *trueprice* variable) following conclusions on the best deal accross different accommodation types were made:

- Compared to other countries, the average prices in Slovakia do not tend to drastically jump.
- On average, the cheapest deals can be found in Poland.
- The optimal distance from the city center is 7,5 miles.
- Autumn is the best season for finding advantageous offers.
- "Tourist" is the class with the lowest prices, however, "standard" is not much more expensive.
- Accommodation on weekdays is not only cheaper, but the prices are also not as dispersed as on weekends.