DRAFT

Alima Dzhanybaeva

2022-10-31

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

- **Some value are non-zero but include zeros. The next table checks the zero values and the quantity of unique values.**

```
kable(funModeling::df_status(df))
```

```
##                  variable q_zeros p_zeros  q_na p_na q_inf p_inf      type
## 1      addresscountryname       0    0.00     0 0.00     0     0 character
## 2             city_actual       0    0.00     0 0.00     0     0 character
## 3      rating_reviewcount       0    0.00 10587 7.06     0     0   integer
## 4         center1distance       0    0.00     0 0.00     0     0 character
## 5            center1label       0    0.00     0 0.00     0     0 character
## 6         center2distance       0    0.00     0 0.00     0     0 character
## 7            center2label       0    0.00     0 0.00     0     0 character
## 8           neighbourhood       0    0.00     0 0.00     0     0 character
## 9                   price       0    0.00     0 0.00     0     0   integer
## 10            price_night       0    0.00     0 0.00     0     0 character
## 11                 s_city       0    0.00     0 0.00     0     0 character
## 12             starrating   26761   17.84     0 0.00     0     0   numeric
## 13             rating2_ta       0    0.00 13037 8.69     0     0   numeric
## 14 rating2_ta_reviewcount      20    0.01 13037 8.69     0     0   integer
## 15      accommodationtype       0    0.00     0 0.00     0     0 character
## 16     guestreviewsrating       0    0.00 10587 7.06     0     0 character
## 17            scarce_room   70944   47.31     0 0.00     0     0   integer
## 18               hotel_id       0    0.00   114 0.08     0     0   integer
## 19                  offer   62346   41.57     0 0.00     0     0   integer
## 20              offer_cat       0    0.00     0 0.00     0     0 character
## 21                   year       0    0.00     0 0.00     0     0   integer
## 22                  month       0    0.00     0 0.00     0     0   integer
## 23                weekend   50164   33.45     0 0.00     0     0   integer
## 24                holiday  117272   78.20     0 0.00     0     0   integer
##    unique
## 1      33
## 2     760
## 3    1083
## 4     143
## 5       1
## 6     151
## 7      48
## 8    1262
## 9    1815
## 10      2
## 11     47
## 12     10
## 13      9
## 14   2315
## 15     23
## 16     29
## 17      2
## 18  22902
## 19      2
## 20      5
## 21      2
## 22      8
## 23      2
## 24      2
```

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
addresscountryname	0	0.00	0	0.00	0	0	character	33
city_actual	0	0.00	0	0.00	0	0	character	760
rating_reviewcount	0	0.00	10587	7.06	0	0	integer	1083
center1distance	0	0.00	0	0.00	0	0	character	143
center1label	0	0.00	0	0.00	0	0	character	1
center2distance	0	0.00	0	0.00	0	0	character	151
center2label	0	0.00	0	0.00	0	0	character	48
neighbourhood	0	0.00	0	0.00	0	0	character	1262
price	0	0.00	0	0.00	0	0	integer	1815
price_night	0	0.00	0	0.00	0	0	character	2
s_city	0	0.00	0	0.00	0	0	character	47
starrating	26761	17.84	0	0.00	0	0	numeric	10

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
rating2_ta	0	0.00	13037	8.69	0	0	numeric	9
rating2_ta_reviewcount	20	0.01	13037	8.69	0	0	integer	2315
accommodationtype	0	0.00	0	0.00	0	0	character	23
guestreviewsrating	0	0.00	10587	7.06	0	0	character	29
scarce_room	70944	47.31	0	0.00	0	0	integer	2
hotel_id	0	0.00	114	0.08	0	0	integer	22902
offer	62346	41.57	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	50164	33.45	0	0.00	0	0	integer	2
holiday	117272	78.20	0	0.00	0	0	integer	2

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'. The new variable names are *distance* and *actualrating* respectively.
  - The original *accommodationtype* variable was tranformed 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 3: Data summary

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

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

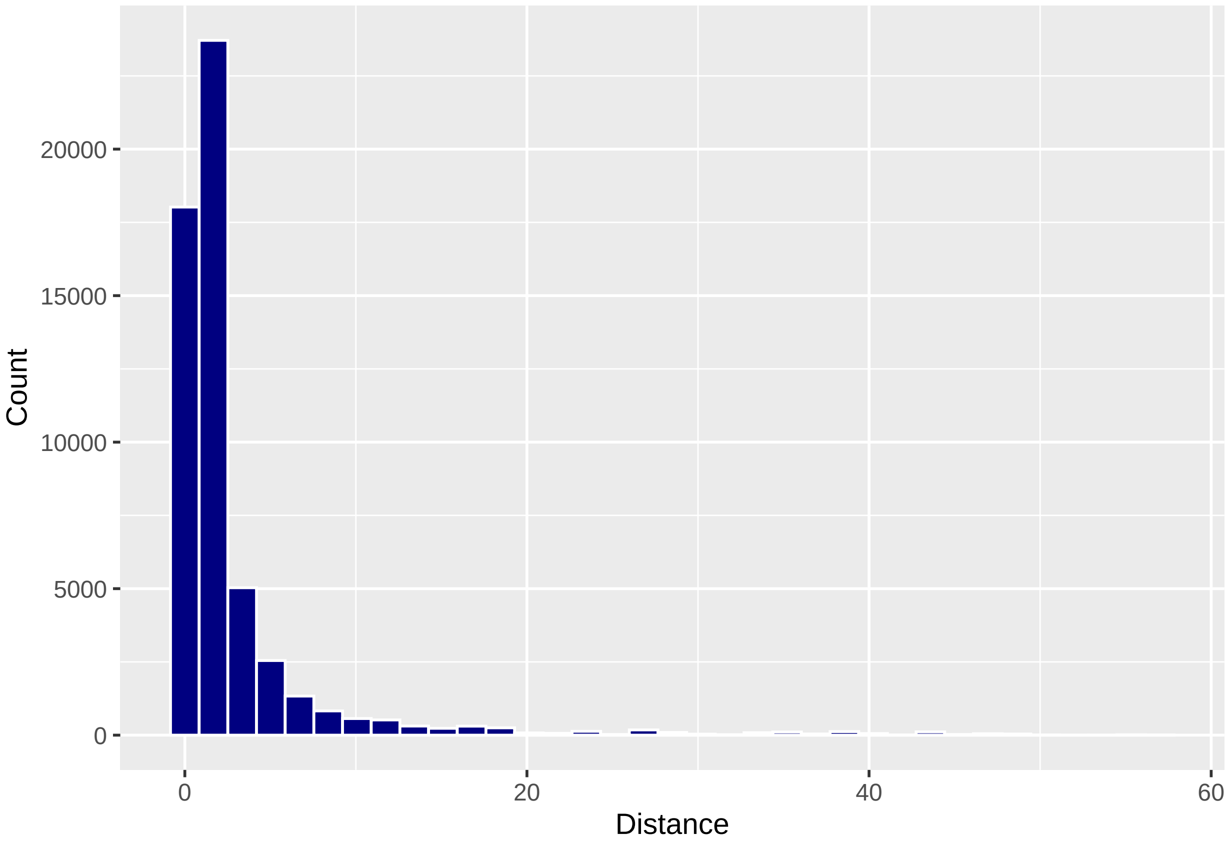
Numeric variables

	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

- **Distance**

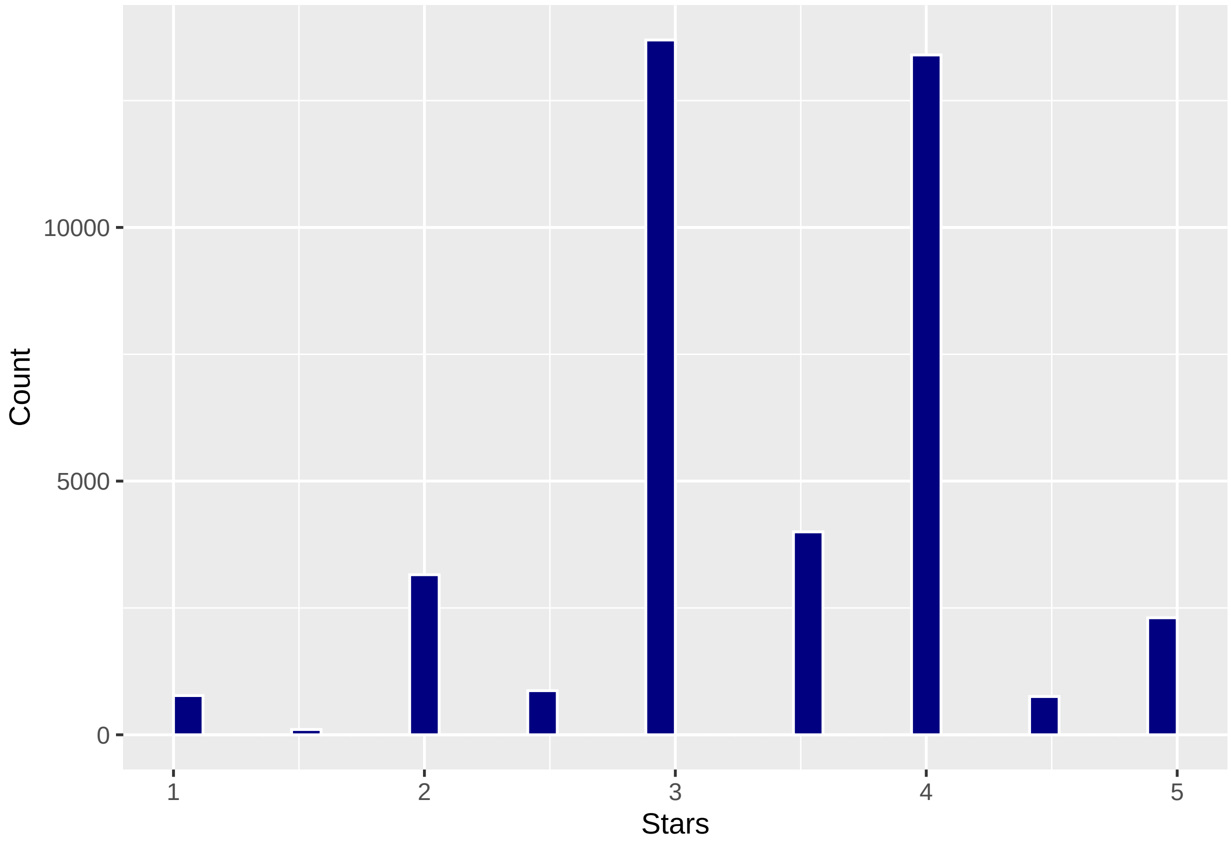
75% of the hotels are located between 0 and 2.4 miles from the main city center.

The histogram below shows the distribution of the distances from the city center, the limit for x-axis was set at 25 miles.



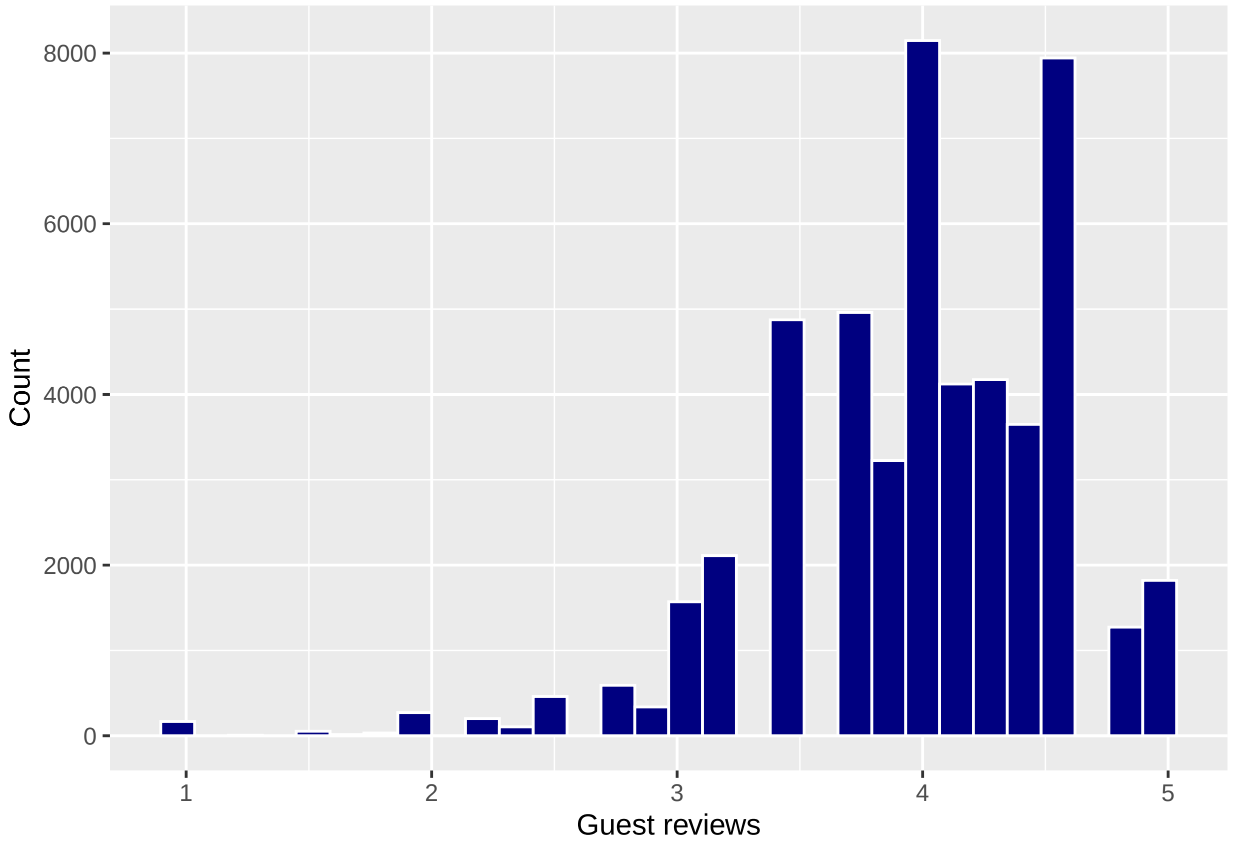
- **Starrating**

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



- **Actualrating**

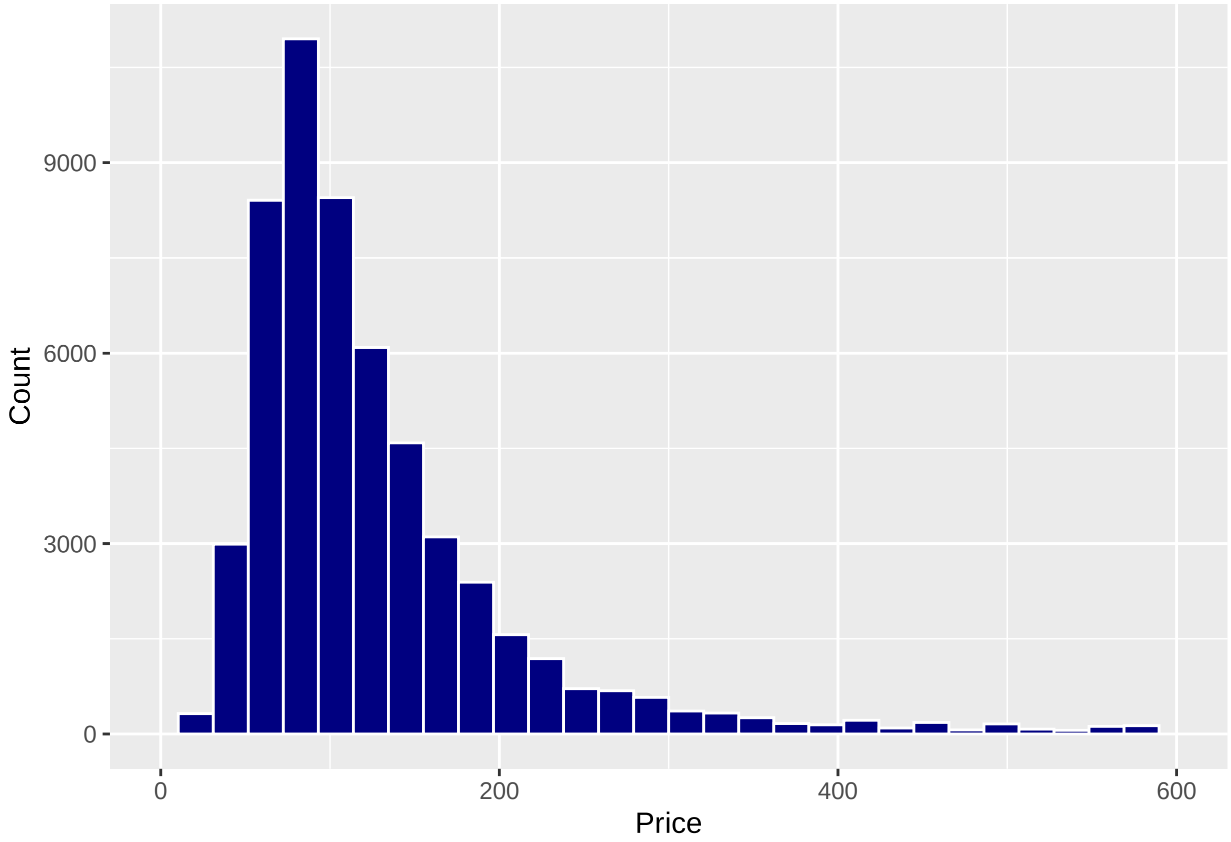
As we can see from the histogram below, 80% of the guest reviews are between 3.4 and 4.1.



- **Trueprice**

99% of the prices are below 600 EUR, thus, the limit for x-axis in histogram was set to this value.

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**

22636 records from the dataset are for the year 2017 and 32311 - for 2018.

year	n
2017	22636
2018	32311

month	n
1	4197
2	6125
3	5340
4	5367
5	5591
6	5691
11	10908
12	11728

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	29880
Apartment	8878
Guest House	7311
Bed and breakfast	4814
Hostel	1618
Apart-hotel	820
Pension	598
Inn	562
Vacation home Condo	335
	31
Caravan Park	26
Motel	25
Country House	18
House boat	17
Chalet	11
Cottage	3

- **Country**

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

```
kable(count(df2, country) %>% arrange(desc(n))) %>%
  kableExtra::kable_styling(position = "center", latex_options = "hold_position")
```

country	n
Italy	27208
Germany	9160
Poland	5780
Austria	5350
Czech Republic	3882
Hungary	2720
Slovakia	847

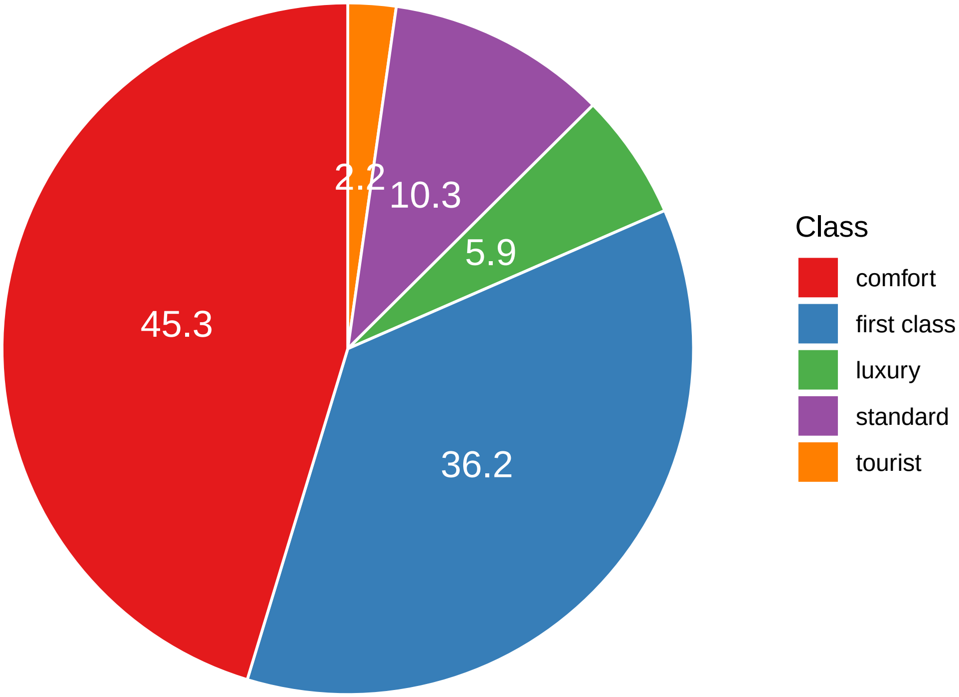
- **Season and class**

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	22050
spring	16298
autumn	10908
summer	5691

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	17695
first class	14154
luxury	2307
standard	4025
tourist	877
NA	15889



Binary variables

- **Weekend and holiday**

36193 records were made on weekends (66%), and 18754 - on weekdays (34%)

weekend	n
0	18754
1	36193

Only 21% (11728) of all observations were made on holidays.

holiday	n
0	43219
1	11728

1.3 Findings.

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

Consequently, we formulated the question:

- *How does the average price change for different values of explanatory variables?*

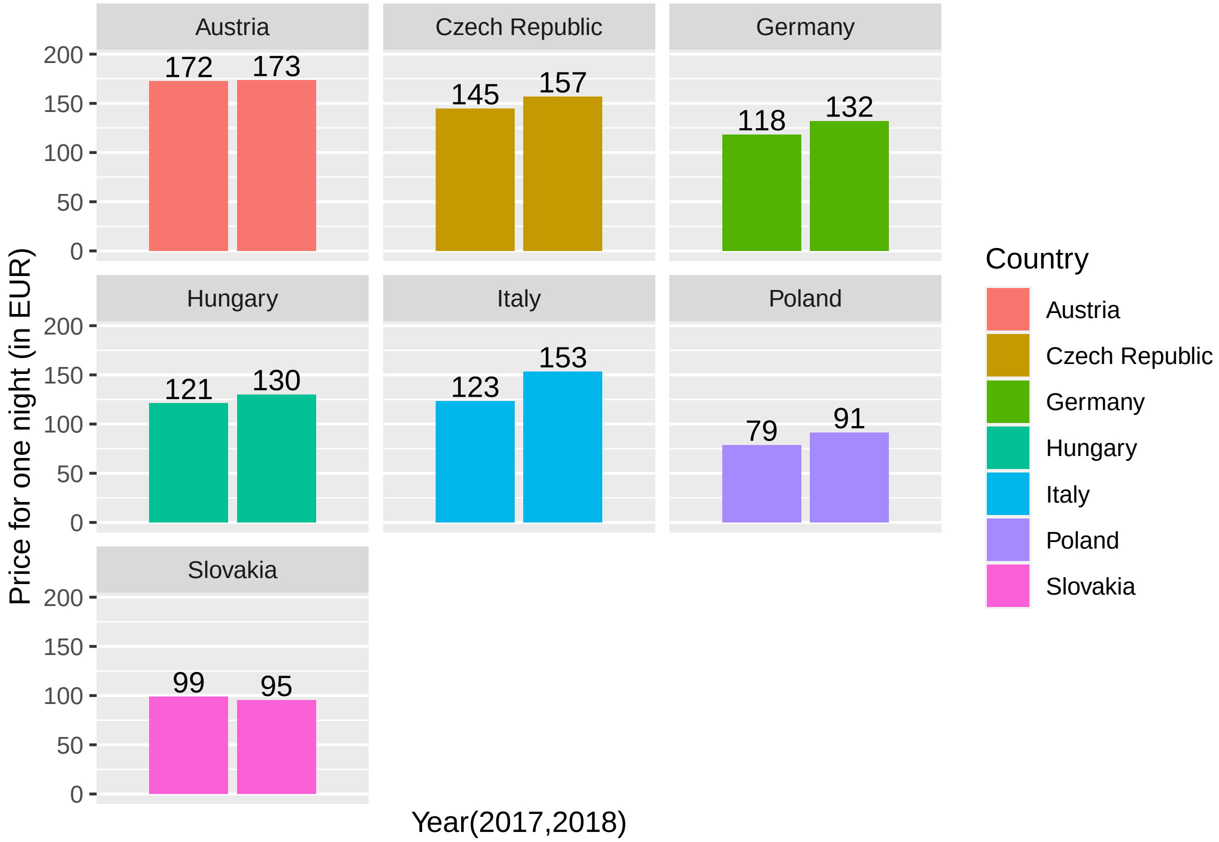
- **Mean prices for countries**

```r
datasummary(trueprice*country ~ Mean + Max + Min, data=df2)%>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

	country	Mean	Max	Min
trueprice	Austria	173.04	6510.00	27.00
	Czech Republic	151.60	7674.00	12.00
	Germany	126.13	1551.00	22.00
	Hungary	126.45	1602.00	19.00
	Italy	140.96	2224.00	16.00
	Poland	86.36	1499.00	16.00
	Slovakia	96.62	1101.00	20.00

- **Plot for diff years diff countries**

```r
df3 <- aggregate(trueprice ~ country + year,
                 data=df2,
                 function(x) {
                   c(mean_price = mean(x))
                 })
ggplot(data = df3, aes(x=year, y=trueprice, fill=country)) +
  geom_bar(stat='identity') + facet_wrap(~country) +
  geom_text(aes(label = round(trueprice)), vjust = -0.2) +
  scale_x_discrete() + ylim(0, 195) +
  labs(x='Year(2017,2018)', y='Price for one night (in EUR)', fill='Country')
```

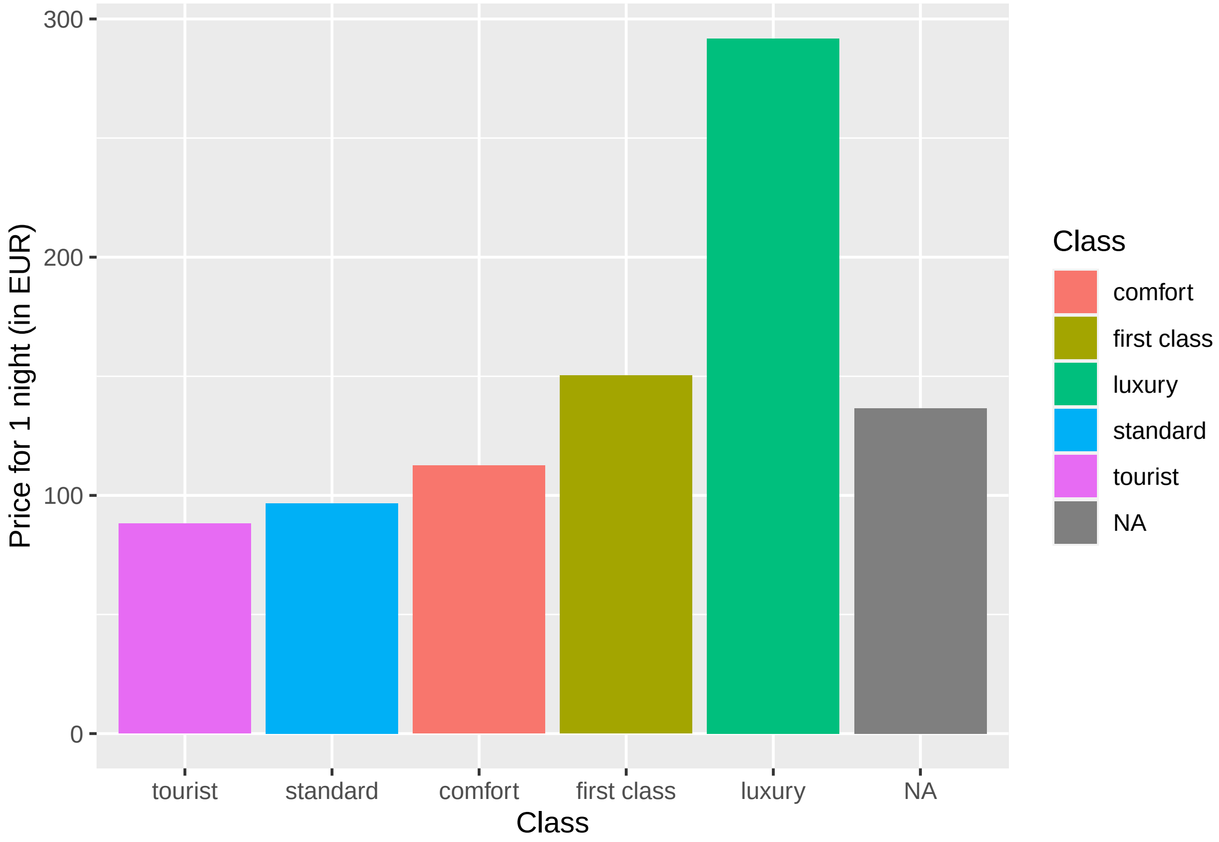


- **Mean prices for diff seasons**

	season	Mean
trueprice	autumn	114.54
	spring	155.01
	summer	151.81
	winter	126.54

- **Plot for classes**

	class	Mean
trueprice	comfort	112.52
	first class	150.36
	luxury	291.85
	standard	96.63
	tourist	88.15


300
200
100
0
Price for 1 night (in EUR)
tourist
standard
comfort
first class
luxury
NA
Class
Class
comfort
first class
luxury
standard
tourist
NA