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 1.1 shows the main variables, their description and type.

```
kable(description, caption = "Table 1.1 Description of the main variables in the original dataset")
```

Table 1: Table 1.1 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.

```
df <- read.csv('https://osf.io/yzntm/download')
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 *guestreviewrating* were changed from 'string' to 'numeric'
- The original *accommodationtype* variable was tranformed into new *accomtype* 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'))
```

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

```
df2 <- df2 %>%
  mutate(class = case_when(starrating == 0.0 | starrating == 1.0 ~ 'tourist',
                           starrating == 1.5 | starrating == 2.0 ~ 'standard',
                           starrating == 2.5 | starrating == 3.0 ~ 'comfort',
                           starrating == 3.5 | starrating == 4.0 ~ 'first class',
                           starrating == 4.5 | starrating == 5.0 ~ 'luxury'))
```

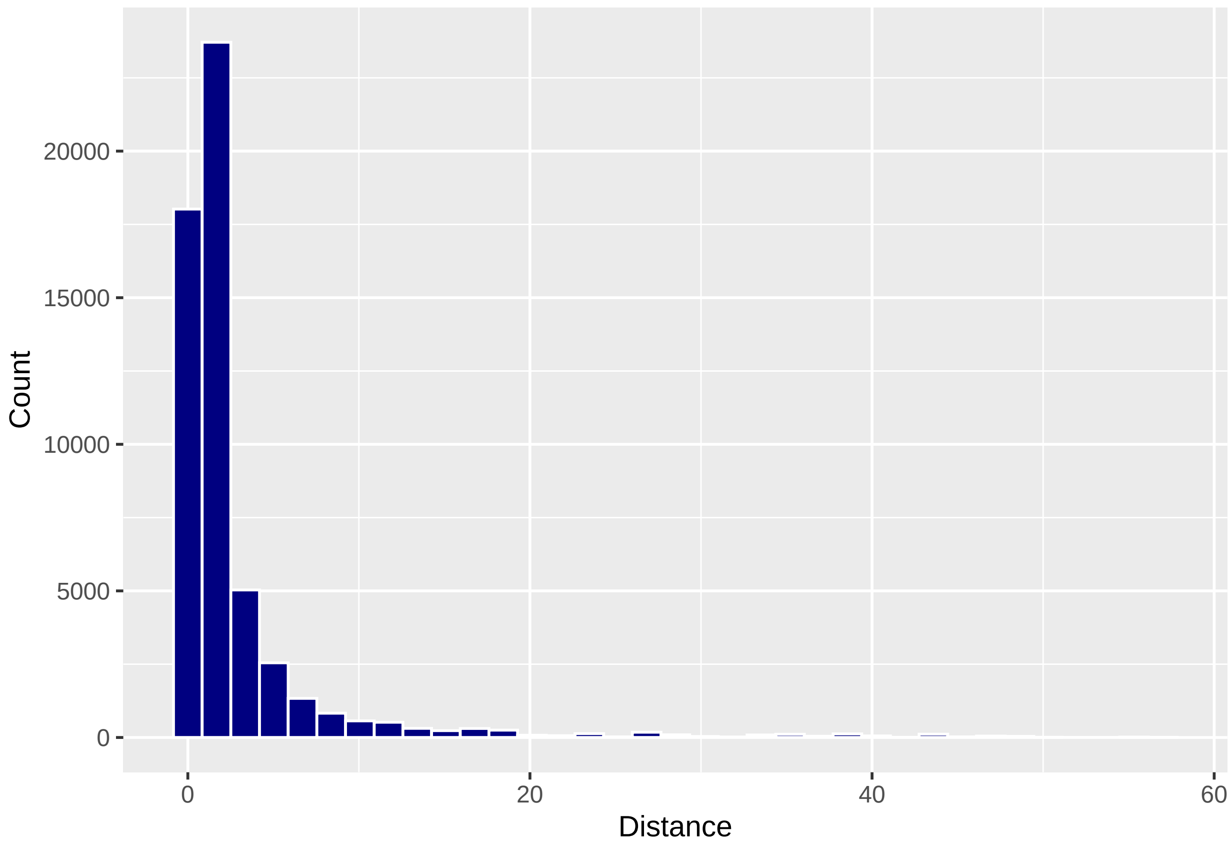
Summary of the transformed dataset

```
datasummary(distance + starrating + actualrating + trueprice ~ Mean + Min + Max + SD, data = df2) %>%
   kableExtra::kable_styling(latex_options = "hold_position")
```

Histogram for distances

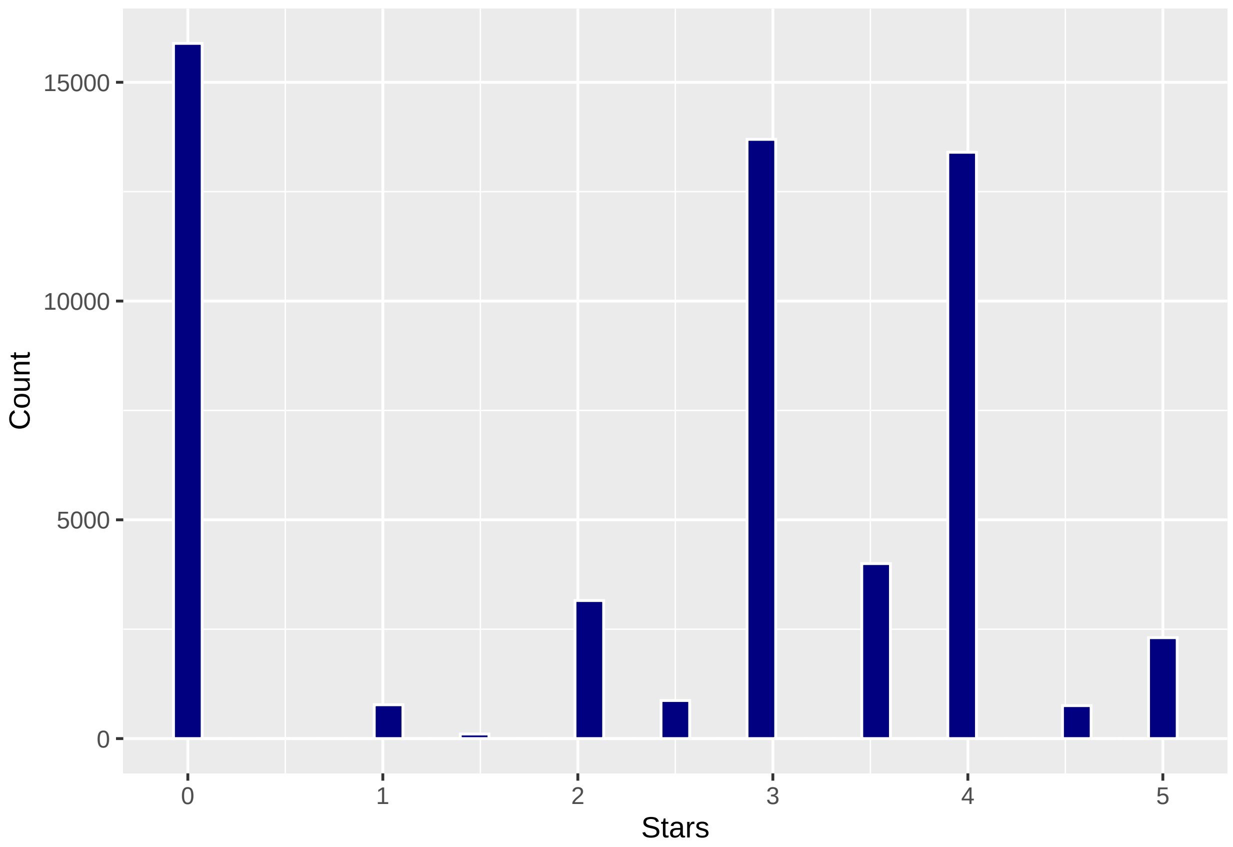
	Mean	Min	Max	SD
distance	2.89	0.00	57.00	5.61
starrating	2.42	0.00	5.00	1.69
actualrating	3.98	1.00	5.00	0.58
trueprice	135.22	12.00	7674.00	129.78

```
ggplot(data=df2, aes(x=distance)) + geom_histogram(fill='navyblue', color='white', bins = 35) +
  labs(x='Distance', y='Count')
```



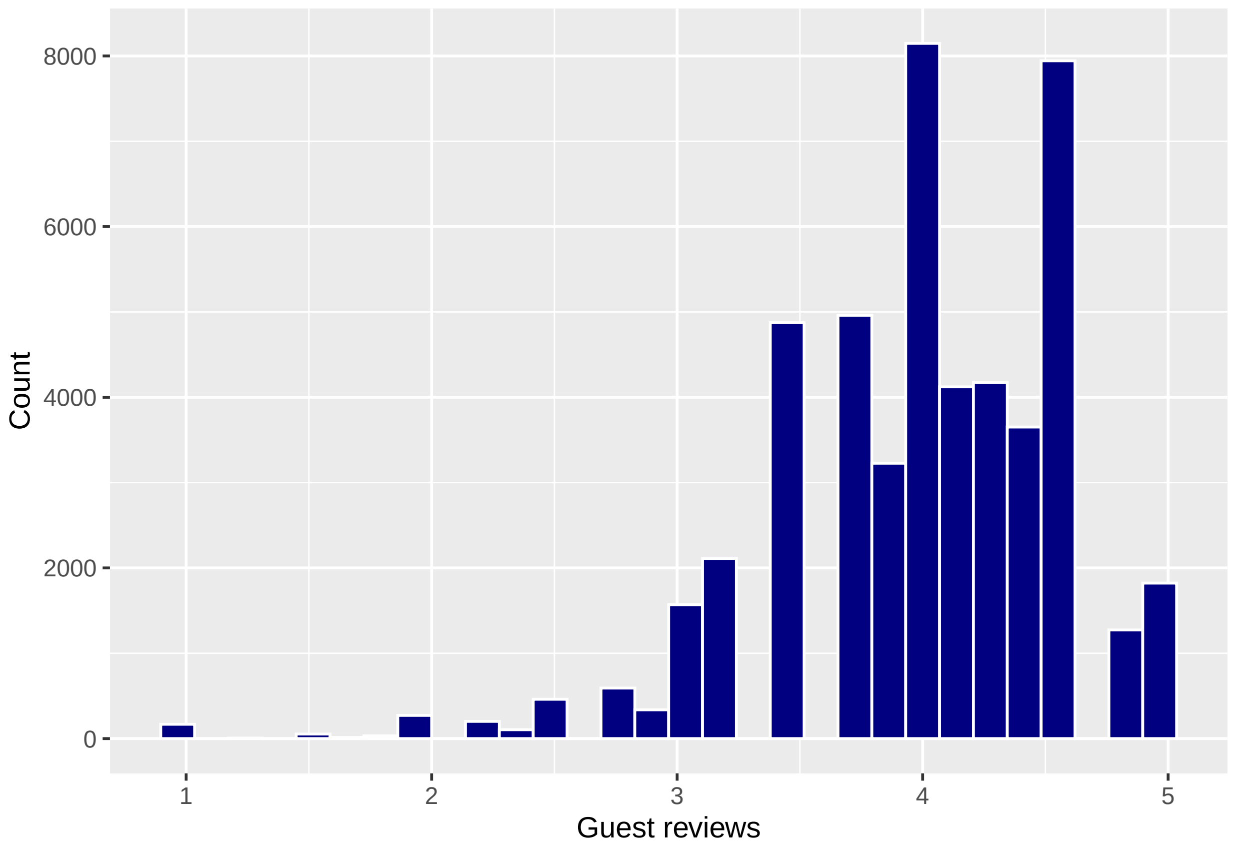
Histogram for starrating

```
ggplot(data=df2, aes(x=starrating)) + geom_histogram(fill='navyblue', color='white', bins = 35) +
  labs(x='Stars', y='Count')
```



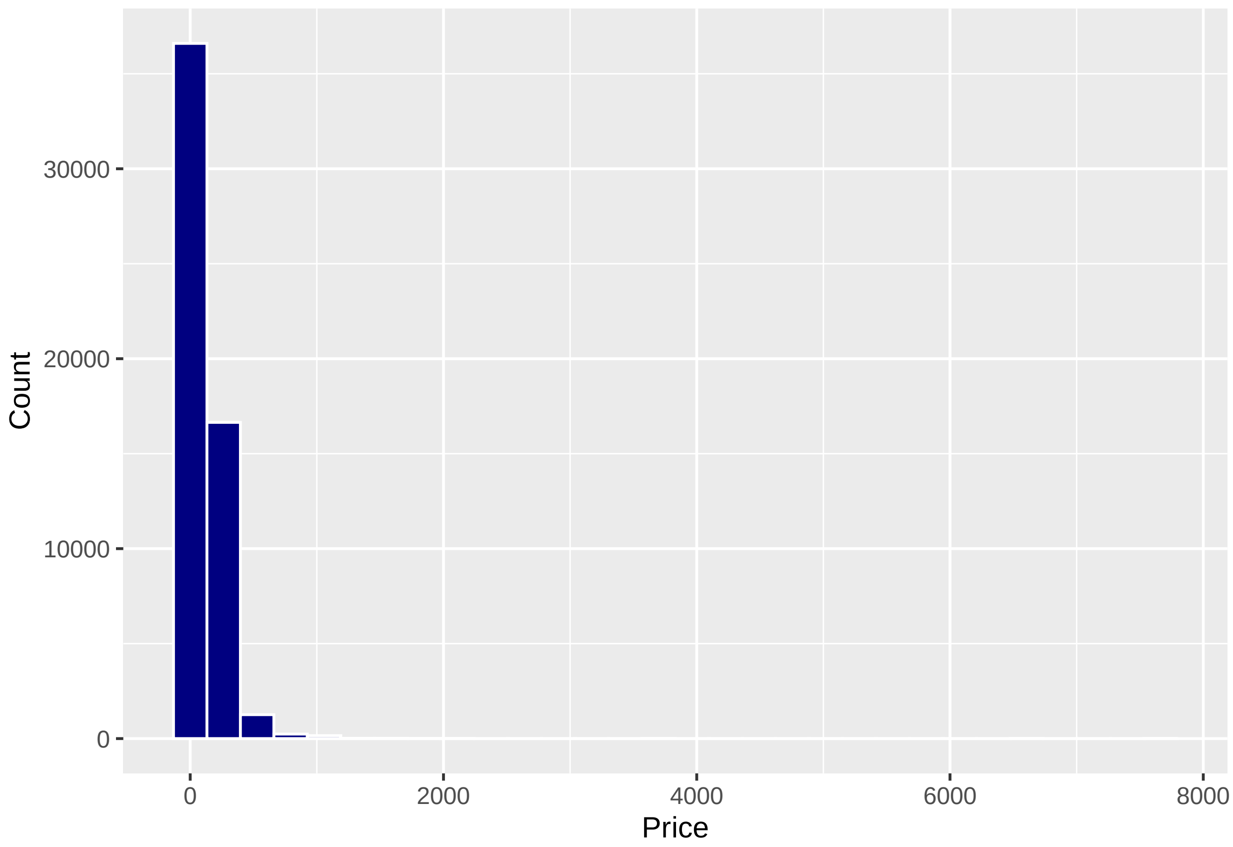
Histogram for guest ratings

```
ggplot(data=df2, aes(x=actualrating)) + geom_histogram(fill='navyblue', color='white') +
  labs(x='Guest reviews', y='Count')
```

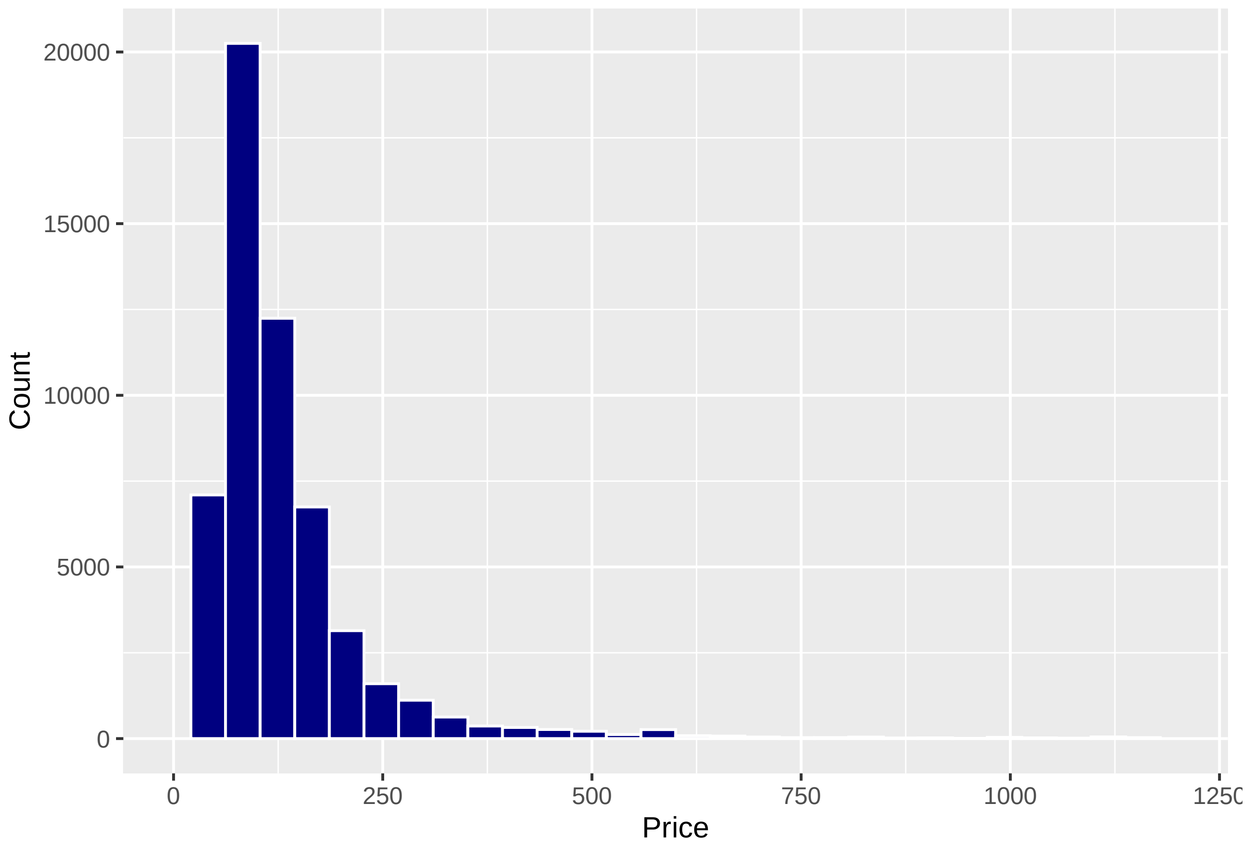


Histogram for prices

```
ggplot(data=df2, aes(x=trueprice)) + geom_histogram(fill='navyblue', color='white') +
  labs(x='Price', y='Count')
```



```
ggplot(data=df2, aes(x=trueprice)) + geom_histogram(fill='navyblue', color='white') +
  labs(x='Price', y='Count') + xlim(0, 1200)
```



Accommodation type

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

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

Weekend and holiday

weekend	n
0	18754
1	36193

holiday	n
0	43219
1	11728

Year and month

year	n
2017	22636
2018	32311

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

Season and class

season	n
winter	22050
spring	16298
autumn	10908
summer	5691

class	n
comfort	14565
first class	17400
luxury	3061
standard	3259
tourist	16662

	addresscountryname	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

	addresscountryname	Mean
actualrating	Austria	4.05
	Czech Republic	4.03
	Germany	3.94
	Hungary	3.97
	Italy	3.95
	Poland	4.04
	Slovakia	4.13

Findings

Mean prices for countries

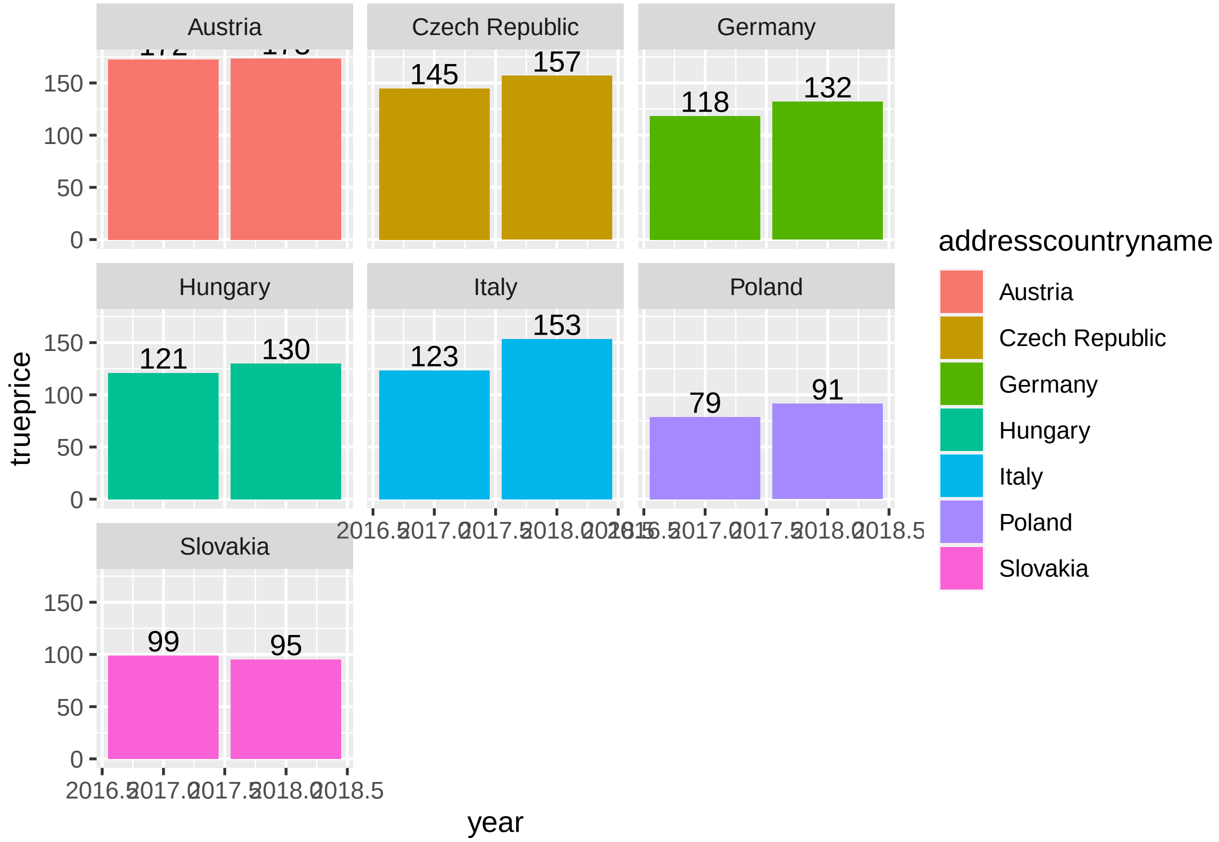
```
datasummary(trueprice*addresscountryname ~ Mean + Max + Min, data=df2)
```

Ratings for diff countries

```
datasummary(actualrating*addresscountryname ~ Mean, data=df2)
```

Plot for diff years diff countries

```
df3 <- aggregate(trueprice ~ addresscountryname + year,
                 data=df2,
                 function(x) {
                   c(mean_price = mean(x))
                 })
ggplot(data = df3, aes(x=year, y=trueprice, fill=addresscountryname)) +
  geom_bar(stat='identity') + facet_wrap(~addresscountryname) +
  geom_text(aes(label = round(trueprice)), vjust = -0.2)
```



Mean prices for diff seasons

```
datasummary(trueprice*season ~ Mean, data=df2) %>%
   kableExtra::kable_styling(latex_options = "hold_position")
```

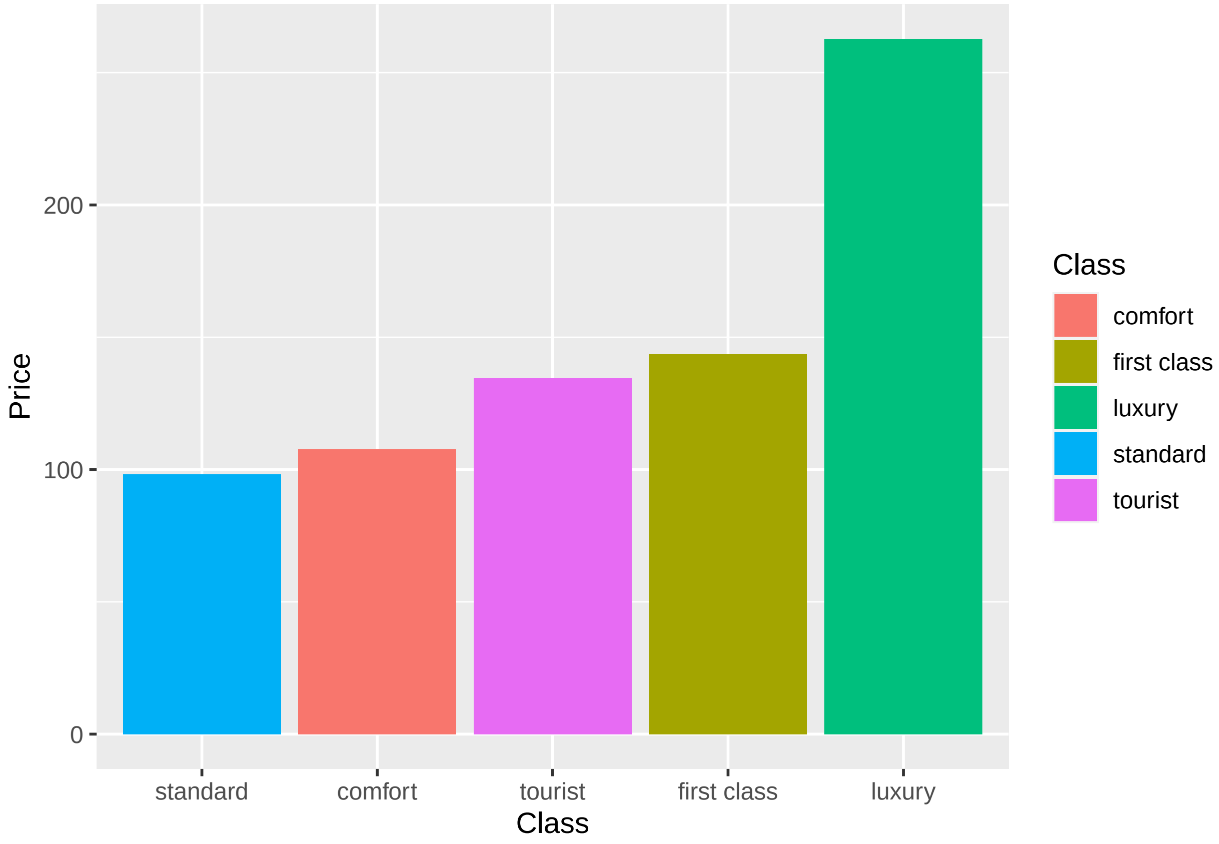
	season	Mean
trueprice	autumn	114.54
	spring	155.01
	summer	151.81
	winter	126.54

Plot for classes

```
datasummary(trueprice*class ~ Mean, data = df2) %>%
   kableExtra::kable_styling(latex_options = "hold_position")
```

```
ggplot(data = df2, aes(x=reorder(class, +trueprice), y=trueprice, fill = class)) +
  geom_bar(stat = 'summary') + labs(x='Class', y='Price', fill = 'Class')
```

	class	Mean
trueprice	comfort	107.65
	first class	143.51
	luxury	262.74
	standard	98.25
	tourist	134.46


Price
standard
comfort
tourist
first class
luxury
Class
Class
comfort
first class
luxury
standard
tourist