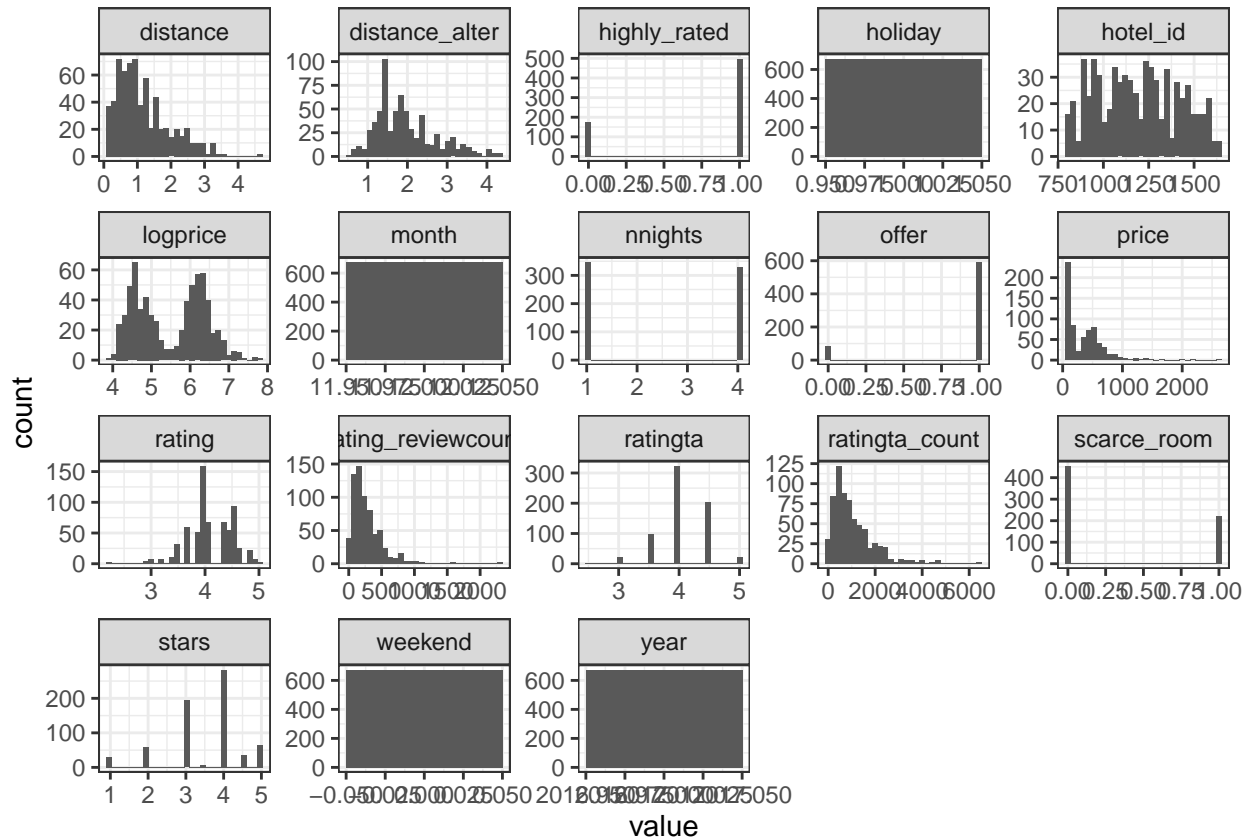

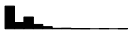





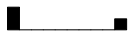












## DA2 Assignment 2 Appendix

### Distributions, Correlations



##	Var1	Var2 value
## 1	hotel_id	hotel_id 1.00
## 2	price	hotel_id -0.07
## 3	offer	hotel_id -0.02
## 8	nnights	hotel_id -0.01
## 9	scarce_room	hotel_id 0.04
## 10	distance	hotel_id 0.15
## 11	stars	hotel_id -0.03
## 12	rating	hotel_id -0.04
## 13	rating_reviewcount	hotel_id -0.08
## 14	ratingta	hotel_id -0.02
## 15	ratingta_count	hotel_id -0.13
## 16	distance_alter	hotel_id 0.07
## 17	logprice	hotel_id -0.06

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
hotel_id	349	0	1189.9	223.9	799.0	1182.0	1632.0	
price	380	0	352.2	305.9	52.0	288.0	2580.0	
offer	2	0	0.9	0.3	0.0	1.0	1.0	
year	1	0	2017.0	0.0	2017.0	2017.0	2017.0	
month	1	0	12.0	0.0	12.0	12.0	12.0	
weekend	1	0	0.0	0.0	0.0	0.0	0.0	
holiday	1	0	1.0	0.0	1.0	1.0	1.0	
nnights	2	0	2.5	1.5	1.0	1.0	4.0	
scarce_room	2	0	0.3	0.5	0.0	0.0	1.0	
distance	35	0	1.2	0.8	0.1	1.0	4.6	
stars	8	0	3.5	1.0	1.0	4.0	5.0	
rating	18	0	4.1	0.4	2.2	4.1	5.0	
rating_reviewcount	267	0	273.0	243.2	3.0	208.5	2303.0	
ratingta	6	0	4.1	0.4	2.5	4.0	5.0	
ratingta_count	317	0	1056.3	901.7	13.0	784.0	6441.0	
distance_alter	36	0	2.0	0.8	0.5	1.8	4.3	
logprice	380	0	5.5	0.9	4.0	5.7	7.9	
highly_rated	2	0	0.7	0.4	0.0	1.0	1.0	

```
## 18      highly_rated      hotel_id 0.02
## 20      price           price 1.00
## 21      offer           price -0.09
## 26      nnights         price 0.75
## 27      scarce_room     price 0.03
## 28      distance        price -0.10
## 29      stars           price 0.28
## 30      rating          price 0.27
## 31 rating_reviewcount   price 0.01
## 32      ratingta        price 0.27
## 33      ratingta_count   price 0.12
## 34      distance_alter   price -0.08
## 35      logprice         price 0.91
## 36      highly_rated     price 0.16
## 39      offer           offer 1.00
## 44      nnights         offer -0.06
## 45      scarce_room     offer 0.01
## 46      distance        offer 0.02
## 47      stars           offer 0.10
## 48      rating          offer 0.05
## 49 rating_reviewcount   offer 0.09
## 50      ratingta        offer 0.08
## 51      ratingta_count   offer 0.10
## 52      distance_alter   offer 0.00
## 53      logprice         offer -0.08
## 54      highly_rated     offer 0.03
## 58      year            year 1.00
```

## 77	month	month	1.00
## 96	weekend	weekend	1.00
## 115	holiday	holiday	1.00
## 134	nnights	nnights	1.00
## 135	scarce_room	nnights	0.03
## 136	distance	nnights	0.01
## 137	stars	nnights	0.02
## 138	rating	nnights	0.02
## 139	rating_reviewcount	nnights	0.01
## 140	ratingta	nnights	0.02
## 141	ratingta_count	nnights	0.01
## 142	distance_alter	nnights	0.01
## 143	logprice	nnights	0.89
## 144	highly_rated	nnights	0.00
## 153	scarce_room	scarce_room	1.00
## 154	distance	scarce_room	-0.11
## 155	stars	scarce_room	-0.19
## 156	rating	scarce_room	-0.03
## 157	rating_reviewcount	scarce_room	-0.20
## 158	ratingta	scarce_room	0.02
## 159	ratingta_count	scarce_room	-0.20
## 160	distance_alter	scarce_room	-0.10
## 161	logprice	scarce_room	0.02
## 162	highly_rated	scarce_room	-0.10
## 172	distance	distance	1.00
## 173	stars	distance	0.17
## 174	rating	distance	-0.03
## 175	rating_reviewcount	distance	-0.19
## 176	ratingta	distance	-0.04
## 177	ratingta_count	distance	-0.04
## 178	distance_alter	distance	0.95
## 179	logprice	distance	-0.09
## 180	highly_rated	distance	0.04
## 191	stars	stars	1.00
## 192	rating	stars	0.55
## 193	rating_reviewcount	stars	0.10
## 194	ratingta	stars	0.46
## 195	ratingta_count	stars	0.26
## 196	distance_alter	stars	0.14
## 197	logprice	stars	0.25
## 198	highly_rated	stars	0.43
## 210	rating	rating	1.00
## 211	rating_reviewcount	rating	0.15
## 212	ratingta	rating	0.84
## 213	ratingta_count	rating	0.26
## 214	distance_alter	rating	-0.02
## 215	logprice	rating	0.25
## 216	highly_rated	rating	0.73
## 229	rating_reviewcount	rating_reviewcount	1.00
## 230	ratingta	rating_reviewcount	0.13
## 231	ratingta_count	rating_reviewcount	0.57
## 232	distance_alter	rating_reviewcount	-0.22
## 233	logprice	rating_reviewcount	0.03
## 234	highly_rated	rating_reviewcount	0.15

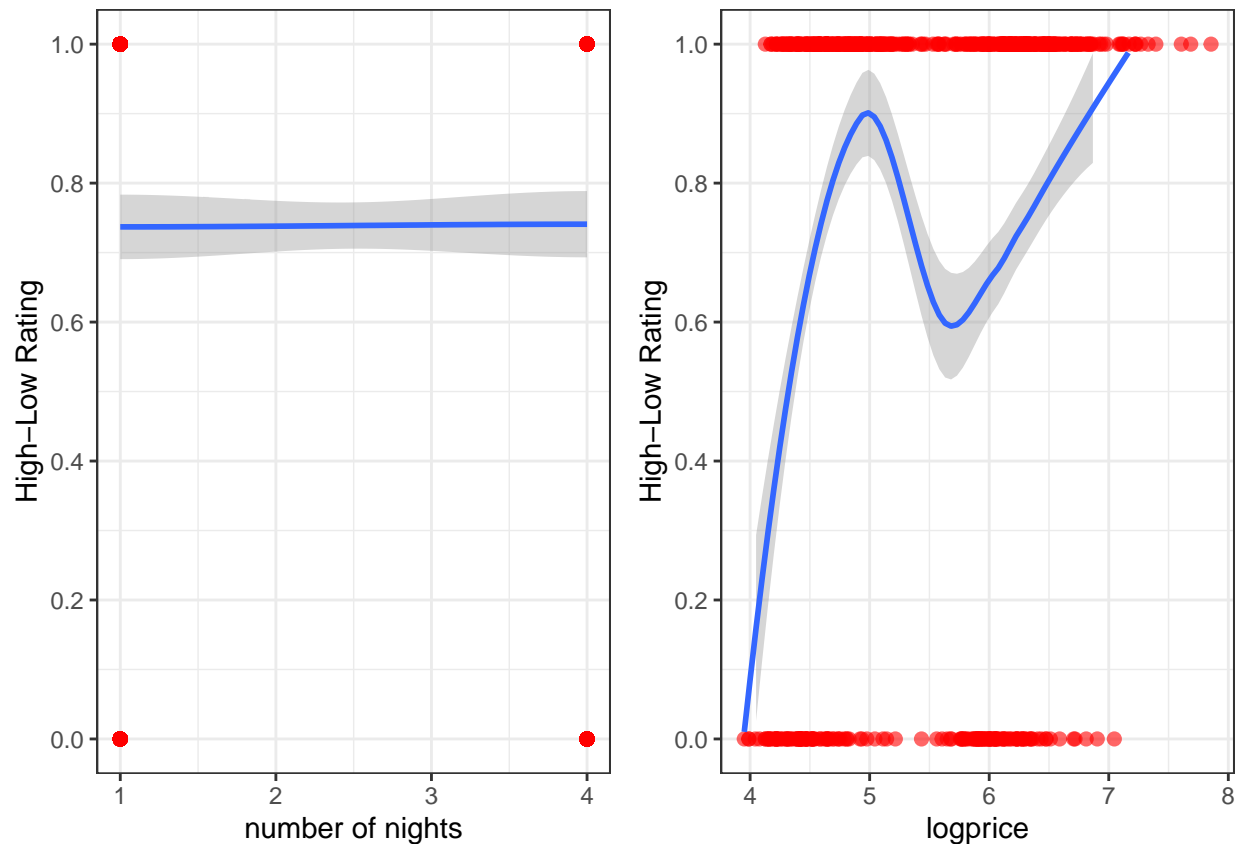
```

## 248          ratingta          ratingta  1.00
## 249  ratingta_count          ratingta  0.25
## 250  distance_alter          ratingta -0.01
## 251          logprice          ratingta  0.25
## 252    highly_rated          ratingta  0.62
## 267  ratingta_count  ratingta_count  1.00
## 268  distance_alter  ratingta_count -0.07
## 269          logprice  ratingta_count  0.12
## 270    highly_rated  ratingta_count  0.23
## 286  distance_alter  distance_alter  1.00
## 287          logprice  distance_alter -0.07
## 288    highly_rated  distance_alter  0.04
## 305          logprice          logprice  1.00
## 306    highly_rated          logprice  0.16
## 324    highly_rated    highly_rated  1.00

```

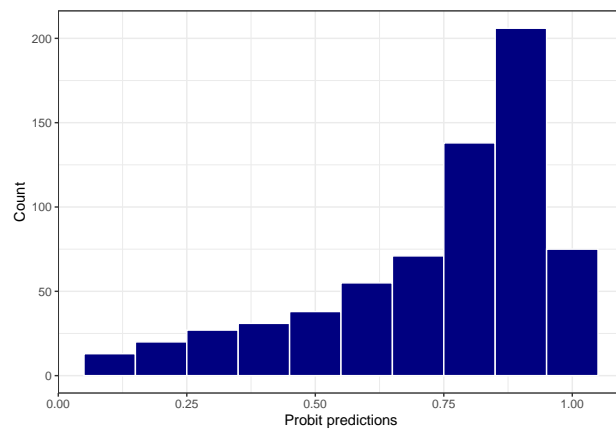
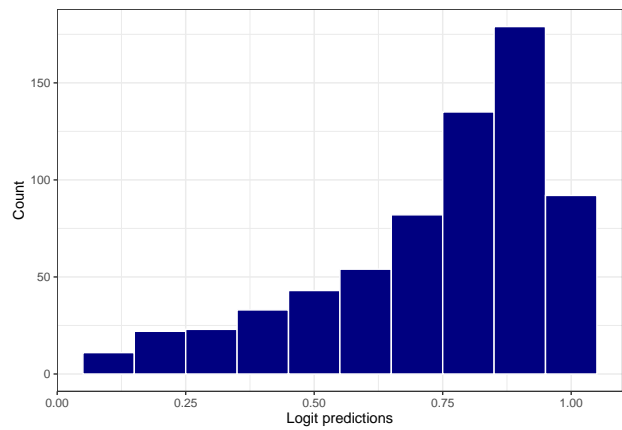
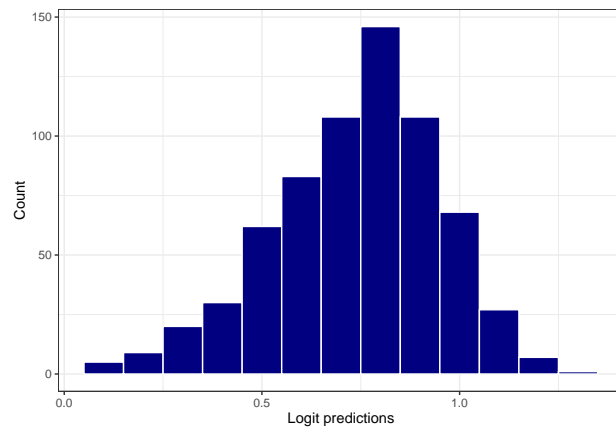
Notably, we can see that prices have lognormal distribution. They are also easier to interpret in % differences. TripAdvisor rating and custome rating have 0.84 correlation. TripAdvisor ratings hopefully largely come from customer ratings so there is likely a big overlap in the two which is why they explain each others variation.

## Lowess graphs



We can see no association for number of nights spent by customer. For logprices, I decided to add knots at  $\ln \text{ price} = 5$  and  $\ln \text{ price} = 5.6$ .

## Histogram of 2 nonlinear models' predictions



As intended nonlinear models don't have predictions above 1.