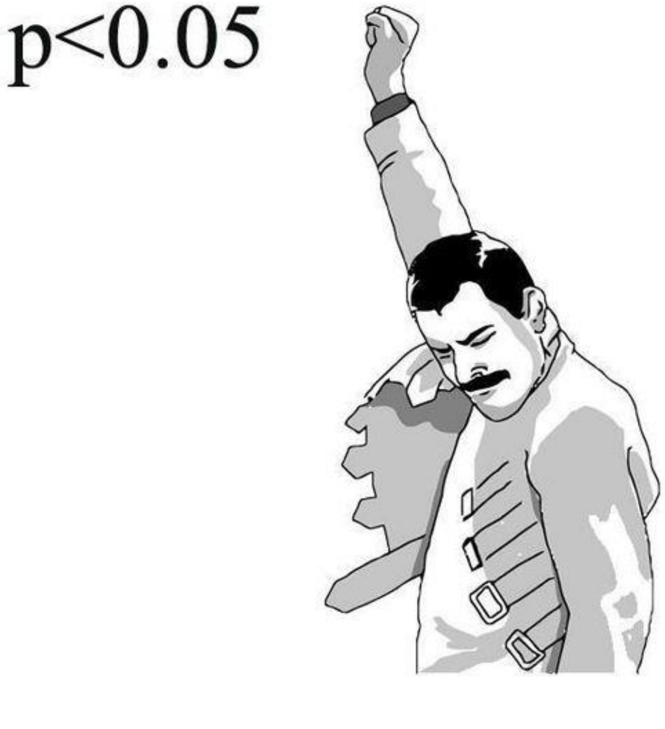
[redacted]

DA3 Assignment #1

Deadline: 23.55h Thursday 16 November 2017



Download hotels_all_nov21.csv. Pick a city. Consider hotels and hostels. Consider all with at least 2 stars.

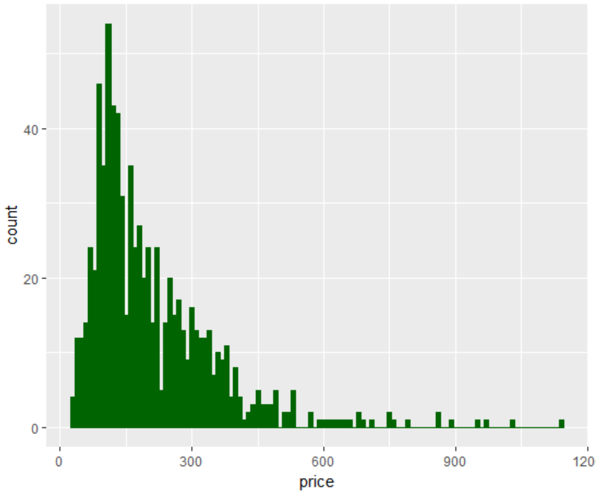
1.Filter the data to the city of your choice and other characteristics (stars, accommodation type). Describe the distribution of the price and distance variables. Comment on graphs. (1-2 sentences)

I picked **London**. There are 1127 accommodations in London, and 792 observations remained after excluding those that are less than 2 stars and not hotels or hostels.

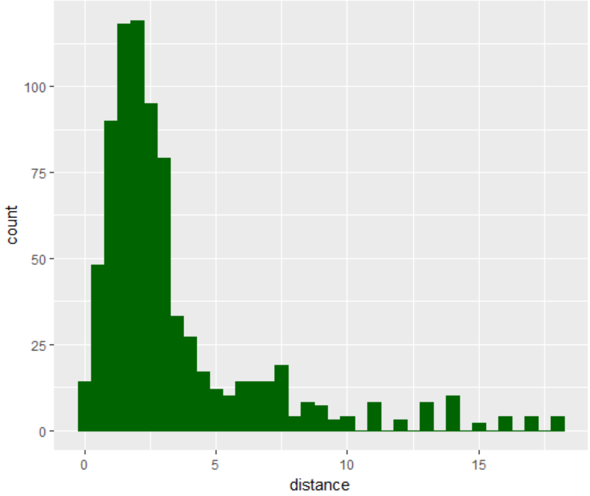
Both the distribution of price and distance are skewed with a long right tail. The prices and the variance of prices are higher if the accommodation is closer to the city center and there are more extreme values in these ranges. There seems to be a negative connection between price and distance (i.e. the higher the distance, the lower the price).

Variable	mean	sd	q.0%	q.25%	q.50%	q.75%	q.100%
price (EUR)	212	148	28	111	167	278	1142
distance (km)	3,4	3,3	0,1	1,4	2,3	3,7	18,0

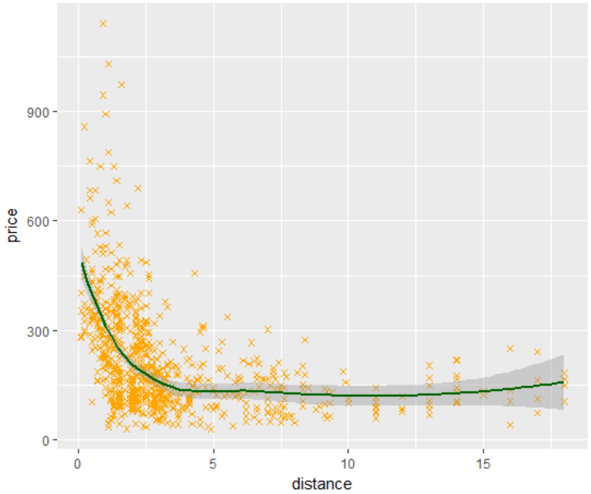
London - price



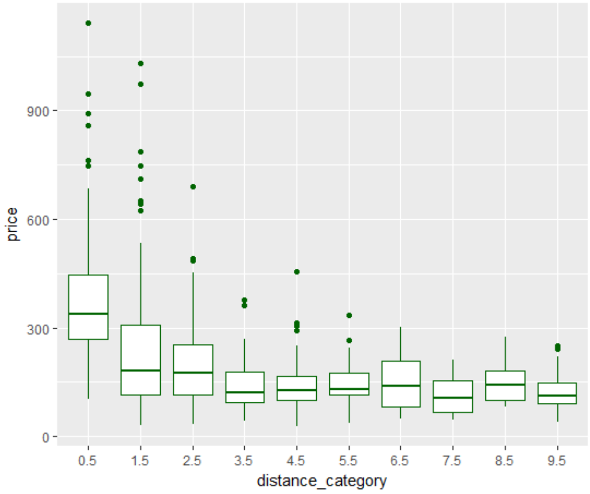
London - distance



London - distance & price



London - distance category & price



2.Sample definition: You may or may not want to drop some observations; make a choice and argue for it (1-2 sentences).

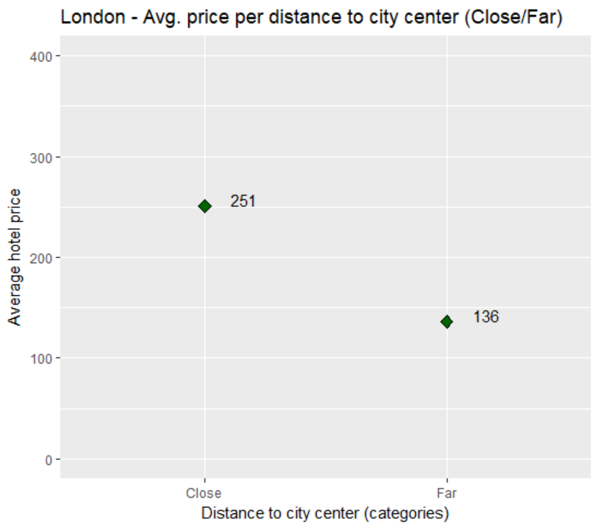
My choice is not to drop any observations. There are extreme values, but these will not influence the regression significantly. There were no duplicated records in the data. There are no observations with missing price or distance values.

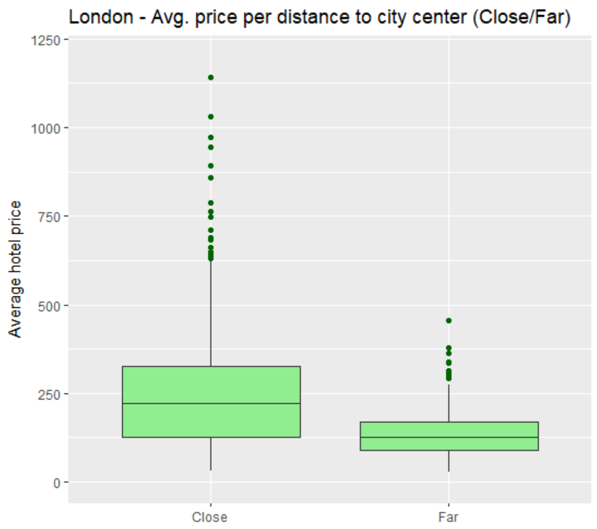
3.Create a binary variable of distance (below/above cutoff of your choice) and regress price on this binary variable. Report, interpret and visualize the results. (1-2 sentences)

There is a break in the strength of the connection at ~ 3 km so the binary variable (dist2) has the value of 'Close' if the accommodation is closer than 3 km ('Far' otherwise).

Average price for hotels that are close is EUR 251, for hotels that are far away is EUR 136. Hotels that are close to the city center are more expensive, on average, than hotels that are further away by EUR 115.

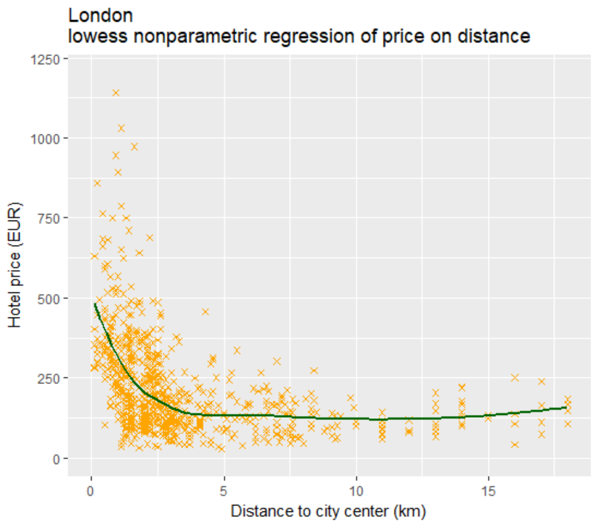
dist2	mean_price	sd_price	min_price	max_price	n
Close	251	163	30	1142	521
Far	136	66	28	457	271





4.Estimate a lowess nonparametric regression of price on distance. Report, interpret and visualize the results. (1-2 sentences)

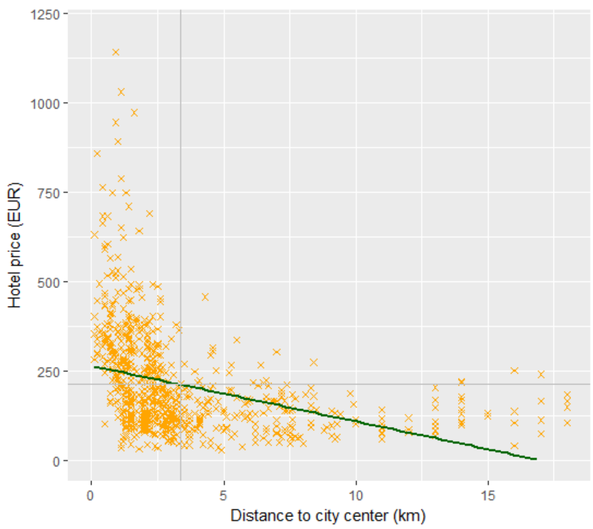
The price – distance relationship is negative in the range of 0 and 3 km. The relationship is basically neutral in the range of 3 and 15 km, and slightly positive above 15 km according to the lowess regression.



5.Estimate a simple linear regression of price on distance. Report, interpret and visualize the results. (1-2 sentences)

Average price is estimated 264 EUR when distance form city center is zero (intercept), and the hotel price is 15.56 EUR lower on average in the hotel that is further away by one km (slope), so the linear regression shows a negative association between price and distance. The regression line goes through the average y, average x point (x = 3.4; y = 212). The regression line is below the observed prices if the distance is greater than 12.

		Intercept				Slope			
Regression	R2	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
01 Level-Level	12%	264,01	7,08	37,27	0,00	-15,56	1,51	-10,32	0,00

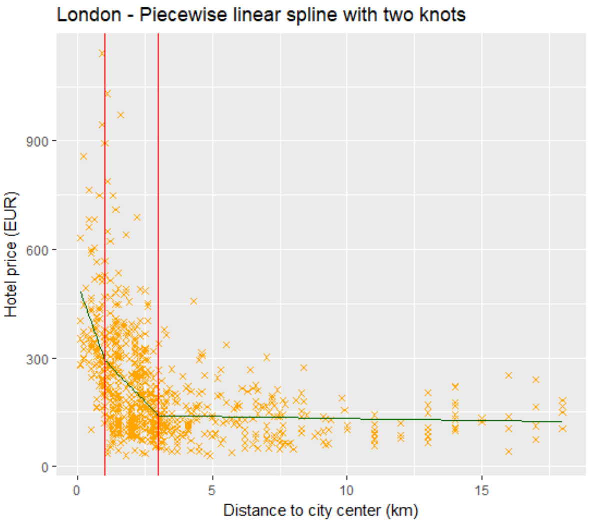


6.Estimate a linear regression of price on distance that captures potential nonlinearities (polynomials, splines). Report, interpret and visualize the results. (1-2 sentences)

I will interpret the **Piecewise linear spline with 2 knots**. The average price is EUR 504.71 when x is zero. Within 1 km of the center the price is expected to be 211 EUR lower for observations with one unit higher distance, and 77 EUR lower for observations with one unit higher distance between 1 and 3 km from the center. If the accommodation is farther than 3 km then the slope does not differ significantly from zero (see table below). R-squared shows that 30% of overall variation of price is captured by the variation predicted by the regression.

Note: I tried to create a piecewise linear spline with 2 knots where the slopes in the different ranges are significant and different from each other but I could not. So I accepted that the slope does differ significantly from zero in some ranges.

		lin. spline 2 knots
	R2	30%
Intercept	Estimate	**504,71**
	Std. Error	28,78
Slope 1	Estimate	**-211,45**
	Std. Error	32,81
Slope 2	Estimate	**-76,69**
	Std. Error	7,26
Slope 3	Estimate	**-1,07**
	Std. Error	1,80
	t value	-0,59
	Pr(>|t|)	0,55



7.Discuss your overall findings. (2-3 sentences)

Looking at the London accommodations we see an obvious negative association between price and distance if the location is close to the city center, but more than 3 km away from the city center this relationship is not significant, the average price does not decrease as we are receding form the center.

I estimated different regressions between price and distance. **Piecewise linear spline with 2 knots** could capture the changes in the slot. The shape of **lowess** is very similar to it. **Simple linear regression** explains only 12% of the variation in the price. **Quadratic** and **Cubic** models reflect non-linearity with higher R-squared but I think that these patterns describe a real life context.

Log-Log model makes sense, since both variables have skewed distribution with long right tail. The intercept of 5.46 means that this is the average ln(price) value when ln(distance) is zero (e.g. when distance is 1 km). The slope of -0,37 means that the price is 0.37% lower on average for observations with one percent increase in the distance. R-squared shows that 25% of overall variation of ln(price) is captured by the variation predicted by the regression.

		Intercept		Slope 1		Slope 2		Slope 3	
Regressions	R2	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Level-Level	12%	264,01	7,08	-15,56	1,51				
Log-Level	13%	5,39	0,03	-0,07	0,01				
Log-Log	25%	5,46	0,03	-0,37	0,02				
Piecewise lspline with 1 knot	29%	409,15	12,38	-92,87	5,81	0,17	1,78		
Piecewise lspline with 2 knots	30%	504,71	28,78	-211,45	32,81	-76,69	7,26	-1,07	1,80
Quadratic	22%	338,92	9,90	-57,24	4,32	2,97	0,29		
Cubic	27%	408,70	13,60	-116,39	9,18	13,33	1,46	-0,44	0,06
Cubic log	27%	5,96	0,06	-0,46	0,04	0,05	0,01	0,00	0,00

+1 For extra point: See what happens when you estimate your models on a selected subsample (i.e. exclude some hotels based on stars, or location.). Discuss the role of cleaning and sample selection.

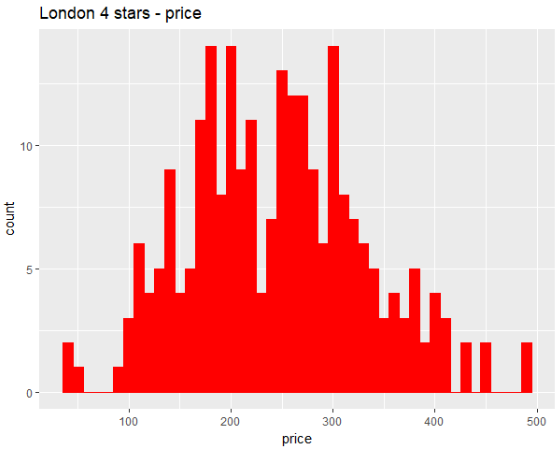
The observations in the selected subsample are the 4 stars London accommodations.

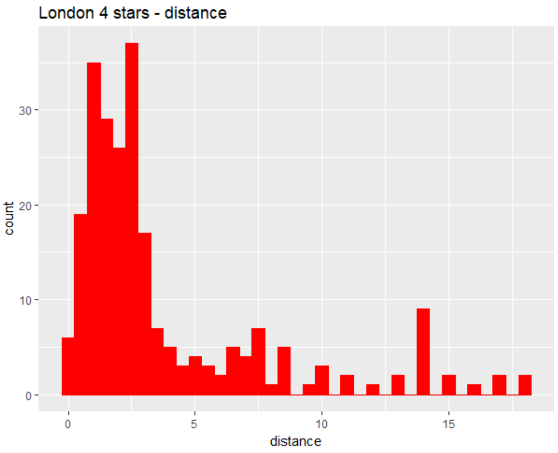
The mean of prices in this sample is higher (212 vs 245), but the standard deviation is lower (148 vs 86). It makes sense since the prices must me more homogeneous in the same stars hotels. The distribution of prices is closer to a normal distribution now.

The mean of distance (3.4 vs 3.7) and the standard deviations (3.3 vs 4.0) did not change that much. Distribution of distance is skewed with a long right tail (similarly to the original sample).

R-squared increased for all regression models. It is not a surprise since the variation of price is smaller compared to the original sample.

Variable	mean	sd	q.0%	q.25%	q.50%	q.75%	q.100%
price (EUR)	245	86	45	182	247	303	491
distance (km)	3,7	4,0	0,1	1,3	2,3	4,1	18,0


London 4 stars - price


London 4 stars - distance

		Intercept		Slope 1		Slope 2		Slope 3	
Regressions – London 4 stars	R2	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Level-Level	30%	289,62	6,38	-11,94	1,17				
Log-Level	29%	5,63	0,03	-0,05	0,01				
Log-Log	34%	5,63	0,03	-0,23	0,02				
Piecewise lspline with 1 knot	42%	356,27	11,07	-49,51	5,42	-5,28	1,42		
Piecewise lspline with 2 knots	45%	326,42	14,57	-21,37	10,57	-93,79	15,33	-3,24	1,55
Quadratic	41%	329,79	8,62	-35,04	3,77	1,55	0,24		
Cubic	42%	345,82	11,82	-49,17	8,10	3,96	1,25	-0,10	0,05
Cubic log	37%	5,84	0,06	-0,18	0,04	0,01	0,01	0,00	0,00