Download cross-country data on life expectancy and GDP per capita. “GDP per capita, PPP (constant)” and “Life expectancy at birth (total)”.

1. Download cross-country data on life expectancy and GDP per capita. “GDP per capita, PPP (constant)” and “Life expectancy at birth (total)”

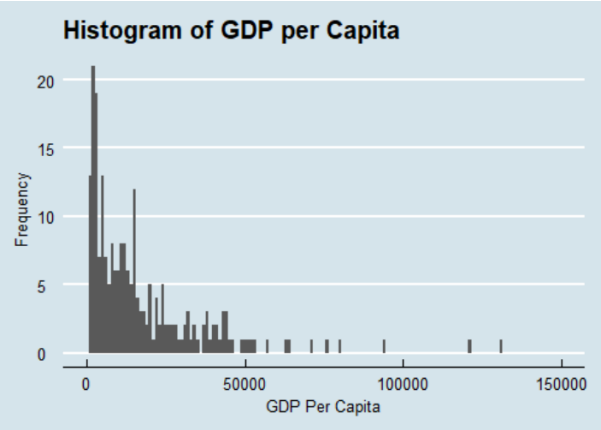
We used API to download the data for 2014. Indicators are NY.GDP.PCAP.PP.KD and SP.DYN.LE00.IN

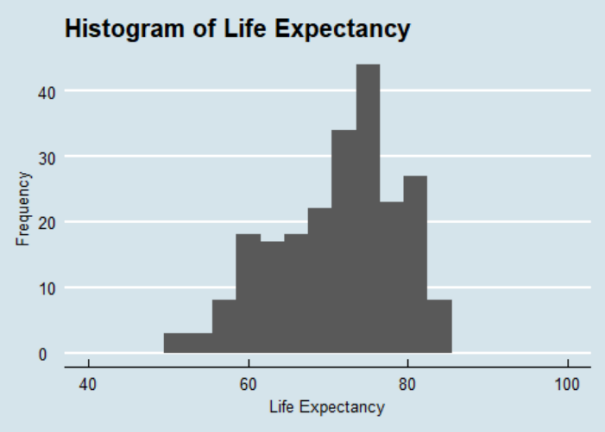
2. Estimate a lowess regression of life expectancy on ln gdp per capita. Estimate a linear regression of life expectancy on GDP per capita that best captures the nonlinearity you found (life expectancy on a piecewise linear spline or a polynomial in the explanatory variable). Argue for your choice. Report the coefficient estimates as well as their confidence interval, interpret and visualize the results.

Descriptive statistics

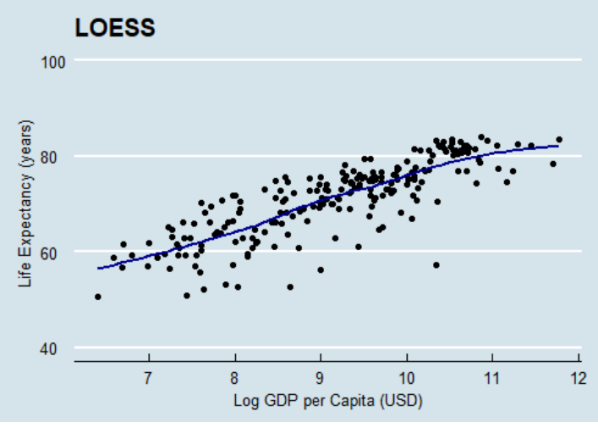
Statistic	N	Mean	St. Dev.	Min	Max
GDP_per_capita	225	17,405	19,497	601.8	130,755.2
life_expectancy	225	71.2	7.8	50.6	84.0

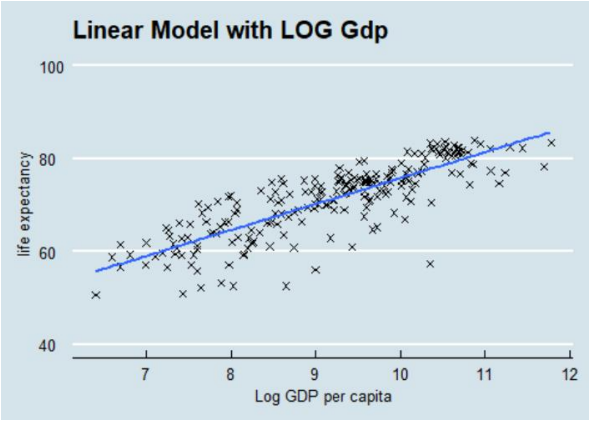
Right skewed distribution for GDP per Capita and sligtly left skewed for Life Expentancy. Taking the log for GDP will make more sense.





Loess is capturing the pattern flexibly but it does not produce parameters which we can interpret and help us generalize.





As we progress and started trying different models, we found the cubic polynomial capturing the variation better than the others. Its R squared was 68.4%. However, its interpretation is much more difficult than the simple linear one. And also its beta2 and beta3 estimates are not significant.

Looking at the results table below we would **pick** the simple linear model with Log GDP per capita. Its R squared is very close to the cubic model and its interpretation is straight forward. Its coefficients are highly significant:

Life expectancy is greater by 5.57 years on average in countries with 10% greater gdp per capita (Number 2 below)

	Dependent variable:			
	life_expectancy			
	(1)	(2)	(3)	(4)
GDP_per_capita	0.0003***			
	(0.00002)			
ln_gdp		5.576***		
		(0.257)		
poly(ln_gdp, 2)1			96.726***	
			(4.452)	
poly(ln_gdp, 2)2			-4.092	
			(4.452)	
poly(ln_gdp, 3)1				96.726***
				(4.438)
poly(ln_gdp, 3)2				-4.092
				(4.438)
poly(ln_gdp, 3)3				-6.806
				(4.438)
Constant	66.688***	19.985***	71.196***	71.196***
	(0.538)	(2.375)	(0.297)	(0.296)
Observations	225	225	225	225
R2	0.415	0.679	0.681	0.684
Adjusted R2	0.412	0.678	0.678	0.680
Note:		*p<0.1; **p<0.05; ***p<0.01		

t test of coefficients for **Model 2**. 95% CI for Beta = [5.07, 6.07]

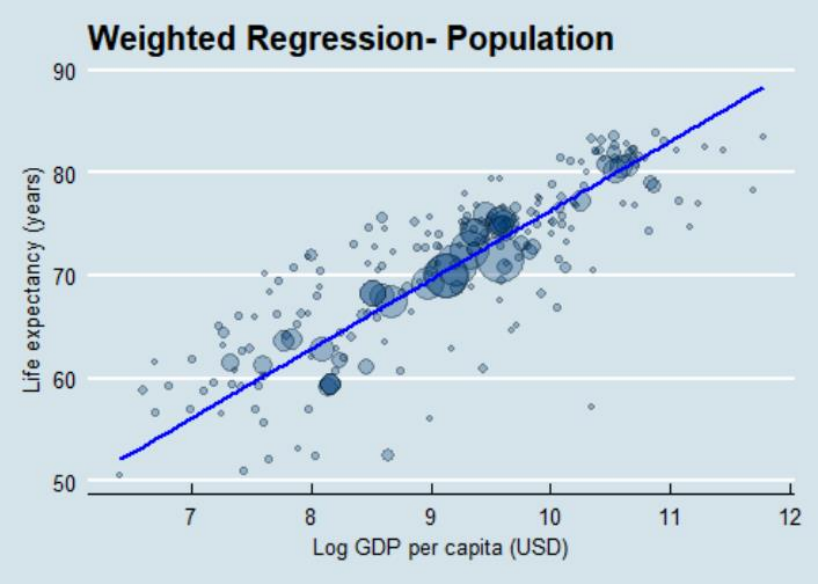
```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 19.98520    2.33999  8.5407 2.098e-15 ***
ln_gdp       5.57617    0.24894 22.4000 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It means that the slope coefficient in the general pattern represented by our data is in the [5.07, 6.07] interval with a 95% chance. We cant see this significance in the cubic model although it has the highest R squared. We will show it below:

t test of coefficients for **Model 4**. 95% CI for Beta2 = [-11.41, 3.23] Zero is in this interval!

```
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)      71.19554    0.29324 242.7885   <2e-16 ***
poly(ln_gdp, 3)1 96.72588    4.19978  23.0312   <2e-16 ***
poly(ln_gdp, 3)2 -4.09188    3.66741  -1.1157   0.2657
poly(ln_gdp, 3)3 -6.80618    4.17966  -1.6284   0.1049
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. **Estimate a weighted regression (weight=population).**




```
================================================
                         Dependent variable:
                     ---------------------------
                            life_expectancy
------------------------------------------------
ln_gdp                         6.729***
                               (0.205)

Constant                       9.037***
                               (1.887)

------------------------------------------------
Observations                     225
R2                               0.829
Adjusted R2                      0.828
Residual Std. Error    39,446.870 (df = 223)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Countries China and India are very close to the regression line and also in the middle of gdp distribution. They dont tilt the regression line much.

People that live in countries with 10% higher GDP per capita live, on average, 0.67 years longer. In this regression we are comparing people living in different countries. In the previous question we were comparing the countries.

Download hotels_all_nov21.csv. Pick a city. Consider hotels and hostels. Consider all with at least 2 stars. You have 6 tasks (1p each).

The goal of the exercise is to use information you have in your data to find a shortlist of five hotels and/or hostels that are good candidates for a good deal. You have to estimate a regression of prices (or log prices) on the other variables of your choice. You have to document your analysis and print the shortlist.

1. Pick a set of variables. Describe all variables used in the analysis.

City: **Milano**.

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
stars	483	3.5	0.7	2.0	5.0
rating	480	3.9	0.4	2.0	4.8
distance	483	5.6	6.7	0.1	30.0
price	483	135	102	39.5	881

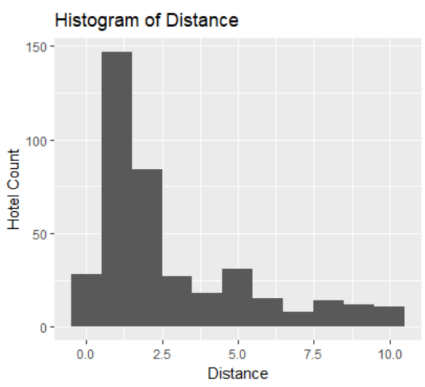
We changed the 3 NA values in rating variable with the average rating for that star category. We are going to focus on 3 variables: Stars, Rating and Distance.

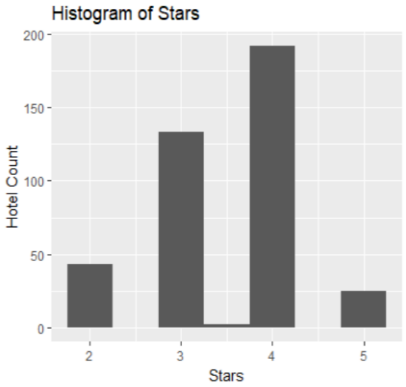
mean_dist	sd_dist	min_dist	max_dist	p50	p95	n
5.58	6.73	0.1	30	2.1	23	483

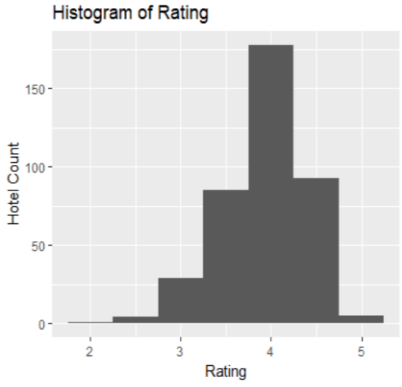
mean_star	sd_star	min_star	max_star	p50	p95	n
3.53	0.73	2	5	4	4.9	483

mean_rate	sd_rate	min_rate	max_rate	p50	p95	n
3.94	0.44	2	4.8	4	4.6	483

We don’t consider it Milano after 10kms. Rating and Stars are slightly left skewed and Distance is heavily right skewed like the price itself. We want to use logs with price and distance during the modelling.







2. Investigate potential nonlinearity of each explanatory variable in simple regressions of the dependent variabl e. Decide on a parametric functional form for each.

We run 3 regressions for each of the explanatory variables: linear, quadratic and cubic.
Regarding nonlinearity, we are observing that the models are performing better when we use cubical form. The variation in price is better captured (explained) in this form. Our parametric functional form choice is cubical form for all three of them. Also their the beta estimates are highly significant. And we have the data for that:

	Dependent variable:		
	lnprice		
	(1)	(2)	(3)
distance	-0.079***		
	(0.009)		
poly(distance, 2)1		-3.990***	
		(0.445)	
poly(distance, 2)2		3.574***	
		(0.445)	
poly(distance, 3)1			-3.990***
			(0.404)
poly(distance, 3)2			3.574***
			(0.404)
poly(distance, 3)3			-3.690***
			(0.404)
Constant	5.037***	4.814***	4.814***
	(0.036)	(0.022)	(0.020)
Observations	395	395	395
R2	0.150	0.270	0.398
Adjusted R2	0.148	0.266	0.393
Residual Std. Error	0.479 (df = 393)	0.445 (df = 392)	0.404 (df = 39)
Note:		*p<0.1; **p<0.05; ***p<0.01	

	Dependent variable:		
	lnprice		
	(1)	(2)	(3)
stars	0.438***		
	(0.026)		
poly(stars, 2)1		6.700***	
		(0.375)	
poly(stars, 2)2		2.500***	
		(0.375)	
poly(stars, 3)1			6.700***
			(0.368)
poly(stars, 3)2			2.500***
			(0.368)
poly(stars, 3)3			1.482***
			(0.368)
Constant	3.279***	4.814***	4.814***
	(0.093)	(0.019)	(0.019)
Observations	395	395	395
R2	0.422	0.481	0.502
Adjusted R2	0.421	0.479	0.498
Residual Std. Error	0.395 (df = 393)	0.375 (df = 392)	0.368 (df =391)
Note:		*p<0.1; **p<0.05; ***p<0.01	

	Dependent variable:		
	lnprice		
	(1)	(2)	(3)
rating	0.610***		
	(0.049)		
poly(rating, 2)1		5.517***	
		(0.427)	
poly(rating, 2)2		2.067***	
		(0.427)	
poly(rating, 3)1			5.517***
			(0.423)
poly(rating, 3)2			2.067***
			(0.423)
poly(rating, 3)3			1.220***
			(0.423)
Constant	2.414***	4.814***	4.814***
	(0.192)	(0.021)	(0.021)
Observations	395	395	395
R2	0.286	0.327	0.341
Adjusted R2	0.285	0.323	0.336
Residual Std. Error	0.439 (df = 393)	0.427 (df = 392)	0.423 (df = 391)
Note:		*p<0.1; **p<0.05; ***p<0.01	

3. Estimate a multiple regression with all explanatory variables in the functional form you specified previously.

```
============================================================
                                Dependent variable:
                            ---------------------------
                                       lnprice
                                  (1)             (2)
------------------------------------------------------------
distance                       -0.080***
                                (0.006)

lndistance                                      -0.279***
                                                 (0.017)

stars                           0.343***        0.313***
                                (0.025)         (0.023)

rating                          0.295***        0.251***
                                (0.043)         (0.039)

Constant                        2.679***        2.914***
                                (0.141)         (0.130)

------------------------------------------------------------
Observations                      395             395
R2                               0.623           0.690
Adjusted R2                      0.620           0.688
Residual Std. Error (df = 391)   0.320           0.290
============================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```

Using logdistance is improving the model performance. Going forward we will use that model which has an R^2 of 69% because it is a better fit and it can capture the variation of price better.
Our beta estimate for lndistance is -0.28. It means price is 0.28% lower in average when we have 1% greater distance keeping all the other variables constant in our data.

And when we compare the hotels with the same distance and rating in this data, price is higher by 31% on average.

4. Pick two slope coefficients, interpret them, and compute and interpret their 95% CI.

We pick the slopes of Stars and Ratings to interpret

```
t test of coefficients:

             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  2.914297   0.135971  21.4332 < 2.2e-16 ***
lndistance  -0.279400   0.017852 -15.6507 < 2.2e-16 ***
stars        0.313201   0.024849  12.6042 < 2.2e-16 ***
rating       0.250843   0.043455   5.7725 1.592e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stars: 95% CI [0.313 - 2x0.025, 0.313 + 2x0.025] = [0.26, 0.36]
Zero is not a part of this interval and the slope is highly significant with a p value even lower than 0.001.

The 95% CI is [0.26, 0.36]. It means that the slope coefficient in the general pattern represented by our data is in the [0.26, 0.36] interval with a 95% chance. That means that, in the general pattern that is represented by our data, average hotel prices are 26 t o 36 percent higher with a 95% chance for hotels that have one more star but are of the same distance from the city center and have the same average customer rating.

Rating: 95% CI [0.250 - 2x0.043, 0.250 + 2x0.043] = [0.16, 0.34]
Zero is not a part of this interval and the slope is highly significant with a p value even lower than 0.001

The 95% CI is [0.16, 0.34].It means that the slope coefficient in the general pattern represented by our data is in the [0.16, 0.34] i nterval with a 95% chance. That means that, in the general pattern that is represented by our data, average hotel prices are 16 t o 34 percent higher with a 95% chance for hotels that have one unit more rating but are of the same distance from the city center and have the same stars.

5. Describe your strategy to find the best deal.

The strategy is to find the lowest residuals. Now all we care about here is prediction, and residuals, in the data. After using our m odel, the hotels with the largest negative residuals are the most underpriced ones compared to how close they are to the city ce nter, what star they have and what rating they got. It is those hotels that we may want to investigate closely. We keep in mind t hat all models are wrong, but some of them are useful.

name	stars	rating	distance	price	lnprice	predMult_2	Mult_2_e
Best Western Hotel Blaise	4	3.9	1.9	65	4.17	4.97	-0.79
Hotel Sempione	3.5	3.9	0.8	73	4.29	5.05	-0.76
Barcelona Milan	4	4.0	4.1	59	4.07	4.78	-0.71
Hotel Milano San Siro	3	4.2	3.1	52	3.96	4.59	-0.64
Idea Hotel Milano SanSiro	4	3.3	5.3	51	3.93	4.53	-0.60

These are fairly close to the city center, have some promising ratings and stars. These hotels definetely deserve a closer look con sidering the point estimate for the Old Town and the overall city price

- Estimate the model of your choice in the previous exercise (ie the exact same dependent variables, same functional for m) for another city of your choice. Discuss your finding.

We picked up Munich, applied the same cleaning, filtering and got estimates from the same functional form.

```
t test of coefficients are highly significant:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.903367   0.193146  9.8546 < 2.2e-16 ***
lndistance  -0.047180   0.011471 -4.1128 6.381e-05 ***
stars        0.309042   0.038847  7.9555 3.802e-13 ***
rating       0.470805   0.045809 10.2776 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
================================================
                        Dependent variable:
                    ---------------------------
                               lnprice
------------------------------------------------
lndistance                    -0.047***
                               (0.014)

stars                          0.309***
                               (0.036)

rating                         0.471***
                               (0.060)

Constant                       1.903***
                               (0.207)

------------------------------------------------
Observations                     156
R2                              0.639
Adjusted R2                     0.631
Residual Std. Error       0.267 (df = 152)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Although R squared reduced a bit, it can still capture 64% of the variation. Beta estimates are highly significant again. Plotting is difficult because we have 3 dimentions.