Exercise 1:

Consider the following dataset with 11 observations and two features: where the last column corresponds to the

	1	2	3	4	5	6	7	8	9	10	11	$\sum_{i=1}^{n}$
y	-7.90	-6.08	-3.74	-1.18	-1.23	-0.55	0.05	0.88	4.74	2.93	2.55	-9.53
x_1	-1.00	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	1.00	0.00
x_2	0.95	0.65	0.40	0.07	0.06	0.02	0.02	0.14	0.34	0.60	0.98	4.23

sum of values of each row.

The following shows the output of an LM ($x_2 \sim x_1$) and a GAM ($x_2 \sim s(x_1)$):

	LM			GAM		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	0.38	0.12 – 0.65	**8.851e-03**	0.38	0.35 – 0.42	**3.196e-07**
x1	-0.01	-0.42 – 0.41	9.749e-01			
s(x1)						**2.542e-05**
Observations	11			11		
R^2 / R^2 adjusted	0.000 / -0.111			0.988		

The R^2-value for the GAM model is the adjusted one.

a) What conclusions could you draw from the LM model output for the relationship between x_1 and x_2?

b) Considering the information provided by the GAM model: How can the previous statement about the relationship between x_1 and x_2 be extended?

Exercise 2:

You are given the bike rental data with the features `season`, `temp`, `hum`, `windspeed`, and `days_since_2011`. A binary target variable y is created:

- Class y=1: "high number of bike rentals" > 70% quantile (i.e., `cnt` > 5531)
- Class y=0: "low to medium number of bike rentals" (i.e., `cnt` $\leq$ 5531)

The following table shows the absolute joint and marginal probabilities of y and `season`.

	WINTER	SPRING	SUMMER	FALL	Σ
y=0	174.00	111.00	98.00	128.00	511.00
y=1	7.00	73.00	90.00	50.00	220.00
Σ	181.00	184.00	188.00	178.00	731.00

a) Calculate and interpret the odds of "high number of bike rentals" vs. "low to medium number of bike rentals" in winter ($\text{odds}_{\text{winter}}$).

b) Calculate and interpret the odds ratio of high vs. low number of bike rentals when `season` changes from winter to spring.

c) Consider the output of a GLM on $y \sim$ `season`:

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-3.2131	0.3854	0.0000
seasonSPRING	2.7941	0.4138	0.0000
seasonSUMMER	3.1280	0.4121	0.0000
seasonFALL	2.2731	0.4199	0.0000

Interpret the β-estimate for the intercept and seasonSPRING.

d) Now compare the two coefficients with the ones in the full model:

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-8.5176	1.2066	0.0000
seasonSPRING	1.7427	0.5977	0.0035
seasonSUMMER	-0.8566	0.7660	0.2635
seasonFALL	-0.6417	0.5543	0.2470
temp	0.2902	0.0391	0.0000
hum	-0.0627	0.0124	0.0000
windspeed	-0.0925	0.0305	0.0024
days_since_2011	0.0166	0.0014	0.0000

Exercise 3:

You are again given the bike sharing data. The target variable `cnt` is renamed in y and the only considered features are `days_since_2011` and `temp`. A linear model with single feature (including intercept) as baselearner (BL) is estimated. The changes in risk (MSE) in each iteration are given in the following tables:

iteration	baselearner	old_risk	new_risk
1	days_since_2011	1 873 827.22	1 733 044.28
2	temp	1733044.28	1 597 057.93
3	days_since_2011	1 597 057.93	1 486 743.19
4	temp	1 486 743.19	1 379 888.98
5	temp	1 379 888.98	1 293 337.07

Calculate the feature importance of the two features.