Exercise 1:

Imagine you work at a second-hand car dealer and are tasked with finding for-sale vehicles your company can acquire at a reasonable price. You decide to address this challenge in a data-driven manner and develop a model that predicts adequate market prices (in EUR) from vehicles' properties.

a) Characterize the task at hand: supervised or unsupervised? Regression or classification? Learning to explain or learning to predict? Justify your answer.

b) How would you set up your data? Name potential features along with their respective data type and state which is the target variable.

c) Assume now that you have data on vehicles' age, mileage, and price. Explicitly define the feature space $\mathcal{X}$ and target space $\mathcal{Y}$ and state the formal notation for an exemplary observation.

d) You choose to use a linear model (LM) for this task. For this, you assume the targets to be conditionally independent given the features, i.e., $y^{(i)}|\mathbf{x}^{(i)} \perp y^{(j)}|\mathbf{x}^{(j)}$ for all $i, j \in \{1, 2, \ldots, n\}, i \neq j$, with sample size n. The linear hypothesis models the target as a linear function of the covariates with Gaussian error term: $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon, \quad \epsilon \sim N(\mathbf{0}, diag(\sigma^2)), \quad \sigma > 0$. Furthermore, you have reason to believe that the effect of mileage might be non-linear, so you decide to include this quantity logarithmically (using the natural logarithm). State the hypothesis space for the corresponding model class. Which parameters need to be learned?

e) Define the corresponding parameter space.

f) State the loss function for the i-th observation using $L2$ loss.

g) In classical statistics, you would estimate the parameters via maximum likelihood estimation (of course, in the special case of the LM, we also have a direct analytical solution via the least-squares estimator). The likelihood for the LM is given by:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^{[t]}\mathbf{x}^{(i)}\right)^2\right) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^{[t]}\mathbf{x}^{(i)}\right)^2\right) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2\right)
\end{aligned}
$$

Describe how you can make use of the likelihood in empirical risk minimization and write down the resulting loss function.

h) Now you need to optimize this risk to find the best parameters and hence the best model via empirical risk minimization. List the necessary steps to solve the optimization problem.

Congratulations, you just designed your first machine learning project!

Exercise 2:

The `mlr3` ecosystem, which we will use heavily to do machine learning in R, mirrors the HRO principle we have encountered in the lecture. Have a look at `https://mlr3book.mlr-org.com/learners.html` for a quick introduction.

a) Familiarize yourself with `tasks`, `learners` and the `train` method. How do these components compare to what you have learned in the lecture?

b) Have a closer look at the learner associated with the hypothesis space of linear models (hint: you can access the learner object's fields and methods with the dollar sign):

```
learner_lm <- mlr3learners::LearnerRegrLM$new()
learner_lm
```

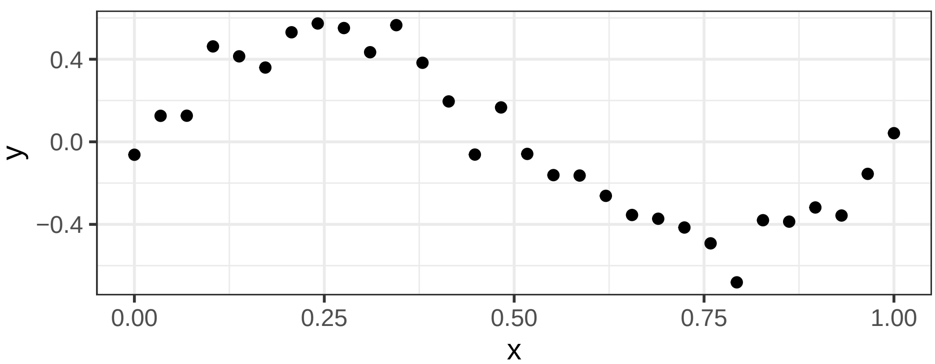
c) We now train a linear regression learner on the `mtcars` data. Describe the task (features, target, number of observations). What does the last line tell you? Lastly, inspect the learner object to get the estimated regression coefficients (hint: you will only find something here after the training method has been called).

```
task_mtcars <- mlr3::tsk("mtcars")
learner_lm$train(task_mtcars)
predictions <- learner_lm$predict(task_mtcars)
predictions$score()
```

Vorschlag: nach Blatt zu Regression schieben

Exercise 3:

Assume the following (noisy) data-generating process: $y = 0.5 + 0.4 \cdot \sin(2\pi x) + \epsilon$ with $\epsilon \sim N(0, 0.1)$.



a) We decide to model the data with a cubic polynomial (including intercept term). State the corresponding hypothesis space.

b) Demonstrate that this hypothesis space is simply a parameterized family of curves by plotting in R curves for 3 different models belonging to the considered model class.

c) State the empirical risk w.r.t. $\boldsymbol{\theta}$ for a member of our hypothesis space. Use $L2$ loss and be as explicit as possible.

d) We can minimize this risk using gradient descent. In order to make this somewhat easier, we will denote our transformed feature matrix enabling us to express our model by $\tilde{\mathbf{X}}\boldsymbol{\theta}$ (note that our model is still linear in its parameters, even if $\mathbf{X}$ has been transformed in a non-linear manner). Derive the gradient of the empirical risk w.r.t. $\boldsymbol{\theta}$, which is multiplied by $-\alpha$ to update the current parameter vector $\boldsymbol{\theta}^{[t]}$.
ggf. mit hints

e) How would the optimization procedure change if we had used a polynomial of different degree (e.g., 1 or 5)?

f) You will not be able to fit the data perfectly with a cubic polynomial. Should you opt for a more flexible model class (i.e., a hypothesis space with higher capacity)? What might be disadvantageous about this?