**Solution 1:**

a) We face a **supervised regression** task: we definitely need labeled training data to infer a relationship between cars' attributes and their prices, and price in EUR is a continuous target (or quasi-continuous, to be exact – as with all other quantities, we can only measure it with finite precision, but the scale is sufficiently fine-grained to assume continuity). **Prediction** is definitely the goal here, however, it might also be interesting to examine the explanatory contribution of each feature.

b) Target variable and potential features:

| **Variable** | **Role** | **Data type** |
|---|---|---|
| Price in EUR | Target | Numeric |
| Age in days | Feature | Numeric |
| Mileage in km | Feature | Numeric |
| Brand | Feature | Categorical |
| Accident-free y/n | Feature | Binary |
| ... | ... | ... |

c) Let $x_1$ and $x_2$ measure age and mileage, respectively. Both features and target are numeric and (quasi-) continuous. It is also reasonable to assume non-negativity for the features, such that we obtain $\mathcal{X} = (\mathbb{R}_0^+)^2$, with $\mathbf{x}^{(i)} = (x_1, x_2)^{(i)} \in \mathcal{X}$ for $i = 1, 2, \dots, n$ observations. As the standard LM does not impose any restrictions on the target, we have $\mathcal{Y} = \mathbb{R}$, though we would probably discard negative predictions in practice.

d) We can write the hypothesis space as:

$$\mathcal{H} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^T\mathbf{x} \mid \boldsymbol{\theta} \in \mathbb{R}^3\} = \{f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \mid (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3\}.$$

Note the **slight abuse of notation** here: in the lecture, we first define $\boldsymbol{\theta}$ to only consist of the feature coefficients, with $\mathbf{x}$ likewise being the plain feature vector. For the sake of simplicity, however, it is more convenient to append the intercept coefficient to the vector of feature coefficients. This does not change our model formulation, but we have to keep in mind that it implicitly entails adding an element 1 at the first position of each feature vector, i.e., $\mathbf{x}^{(i)} := (1, x_1, x_2)^{(i)} \in \{1\} \cup \mathcal{X}$, constituting the familiar column of ones in the design matrix $\mathbf{X}$.

e) The parameter space is included in the definition of the hypothesis space and in this case given by $\Theta = \mathbb{R}^3$.

f) Loss function for the $i$-th observation: $L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) = \left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2$.

g) The first thing to note is that both MLE and ERM are **optimization problems**, and both should lead us to the same optimum. Their opposite signs are not a problem: maximizing the likelihood is equivalent to minimizing the negative likelihood. Also, both are defined pointwise. The last thing to fix is therefore the product introduced by the independence assumption in the joint likelihood of all observations (recall that we use a *summed* loss in ERM), for which the logarithm is a natural remedy. We can thus simply use the **negative log-likelihood (NLL)** as our loss function (and indeed, many known loss functions can be shown to correspond to certain model likelihoods).

Let's put these reflections to practice:

$$
\begin{aligned}
L_{NLL}\left(y^{(i)}, f\left(\mathbf{x}^{(i)}|\boldsymbol{\theta}\right)\right) &= -\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\
&= -\ell(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\
&= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2\right)\right) \\
&= -\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2\right)\right)\right) \\
&= -\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2\right) \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} -\ell(\boldsymbol{\theta}|\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{n} L_{NLL}\left(y^{(i)}, f\left(\mathbf{x}^{(i)}|\boldsymbol{\theta}\right)\right) \\
&= \sum_{i=1}^{n} \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2 \\
&= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2 \\
&\propto \sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2 \\
&= \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) \quad (L2 \text{ loss})
\end{aligned}
$$

As we are only interested in the feature coefficients here, we neglect all irrelevant terms that do not depend on $\boldsymbol{\theta}$ as they have no effect on the solution (i.e., the $\arg\min$ of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$). This is what the proportional sign $\propto$, often used in contexts of optimization and Bayesian statistics, means: we keep only expressions impacted by our parameter of interest because they suffice to yield the intended results or show some property of interest.

From this we can easily see the correspondence between MLE and ERM: the $L2$ loss is proportional to the negative log-likelihood and hence, the $\arg\max$ of the likelihood (using the assumption of normally distributed errors) and the $\arg\min$ of the risk (using $L2$ loss) are equivalent.

h) In order to find the optimal $\hat{\boldsymbol{\theta}}$, we need to solve the following minimization problem:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\theta}\in\Theta}\left(\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2\right) \\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2
\end{aligned}
$$

This is achieved in the usual manner of setting the derivative w.r.t. $\boldsymbol{\theta}$ to 0 and solving for $\boldsymbol{\theta}$, yielding the familiar least-squares estimator $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.