1 Problem 1

a)
we need to show the independence of x and y, we've known the joint density of x and y. we can compute the marginal density of x ($p(x)$) and y ($p(y)$). if $p(x) \times p(y) = p(x, y)$, then we could say that, x, y are independent.
The formula of joint density of x, y:

$$p(x, y) = \lambda\eta e^{-\lambda x - \eta y}, \qquad \lambda, \eta > 0, \qquad x, y \in \mathbb{R}_+^2$$

The marginal density of x:

$$\begin{aligned}
p(x) = \int_0^\infty p(x, y)dy &= \int_0^\infty \lambda\eta e^{-\lambda x - \eta y} dy \\
&= \lambda\eta e^{-\lambda x} \int_0^\infty e^{-\eta y} dy \\
&= \lambda\eta e^{-\lambda x} \int_0^\infty -\frac{1}{\eta} e^{-\eta y} d(-\eta y) \\
&= -\lambda e^{-\lambda x} \times e^{-\eta y} \Big|_0^\infty \\
&= \lambda e^{-\lambda x}
\end{aligned}$$

The marginal density of y:

$$\begin{aligned}
p(y) = \int_0^\infty p(x, y)dx &= \int_0^\infty \lambda\eta e^{-\lambda x - \eta y} dx \\
&= \lambda\eta e^{-\eta y} \int_0^\infty e^{-\lambda x} dx \\
&= \lambda\eta e^{-\eta y} \int_0^\infty -\frac{1}{\lambda} e^{-\lambda x} d(-\lambda x) \\
&= -\eta e^{-\eta y} \times e^{-\lambda x} \Big|_0^\infty \\
&= \eta e^{-\eta y}
\end{aligned}$$

The product of marginal density of x and y:

$$p(x) \times p(y) = \lambda e^{-\lambda x} \times \eta e^{-\eta y} = \lambda\eta e^{-\lambda x - \eta y} = p(x, y)$$

Proved, x, y are independent.

b)
We know that x_i and y_i in D are independent, we also know the joint density function of x, y. Hence, the ML function is on below.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{N} p((x_i, y_i)|\lambda) = \prod_{i=1}^{N} \lambda\eta e^{-\lambda x_i - \eta y_i} \\
l(\lambda) &= \ln L(\lambda) = \sum_{i=1}^{N} \ln\left(\lambda\eta e^{-\lambda x_i - \eta y_i}\right) \\
&= \sum_{i=1}^{N} [\ln\lambda - \lambda x_i + \ln\eta - \eta y_i] \\
&= N(\ln\lambda + \ln\eta) + \sum_{i=1}^{N}(-\lambda x_i - \eta y_i) \\
\frac{\partial l(\lambda)}{\partial\lambda} &= \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0 \\
\hat{\lambda} &= \frac{N}{\sum\limits_{i=1}^{N} x_i}
\end{aligned}
$$

But we need to check whether second derivative is smaller than 0, if yes, we could conclude that $\hat{\lambda}$ is ML estimator.

$$
\frac{\partial l(\lambda)}{\partial\lambda^2} = -\frac{N}{\lambda^2} < 0
$$

Therefore, maximum likelihood estimator of the parameter λ is $\hat{\lambda} = {}^{N}\!/\!\sum_{i=1}^{N} x_i$

c)
under the constraint $\eta = 1/\lambda$, what is the maximum likelihood estimator of the parameter λ?
The process is similar with the upper part.

$$l(\lambda) = \ln L(\lambda) = \sum_{i=1}^{N} \ln\left(e^{-\lambda x_i - \frac{1}{\lambda} y_i}\right)$$

$$= \sum_{i=1}^{N} (-\lambda x_i - \frac{1}{\lambda} y_i)$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = -\sum_{i=1}^{N} x_i + \frac{\sum_{i=1}^{N} y_i}{\lambda^2} = 0$$

$$\lambda^2 = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

$$\hat{\lambda} = \sqrt{\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i}$$

check the second derivative:

$$\frac{\partial l(\lambda)}{\partial \lambda^2} = -2 \sum_{i=1}^{N} y_i / \lambda^3 < 0$$

Therefore, maximum likelihood estimator of the parameter λ is $\hat{\lambda} = \sqrt{\sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i}$

d)
under the constraint $\eta = 1 - \lambda$

$$
\begin{aligned}
l(\lambda) = \ln L(\lambda) = N \ln[\lambda(1-\lambda)] + \sum_{i=1}^{N} \ln \left(e^{-\lambda x_i - \frac{1}{\lambda} y_i} \right) \\
= N \ln(\lambda) + N \ln(1-\lambda) + \sum_{i=1}^{N} (-\lambda x_i + (1-\lambda) y_i) \\
\frac{\partial l(\lambda)}{\partial \lambda} = \frac{N}{\lambda} - \frac{N}{1-\lambda} - \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i = 0 \\
\lambda^2 (\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i) + \lambda (\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i - 2N) + N = 0 \\
\hat{\lambda} = \frac{(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i - 2N) \pm \sqrt{(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i)^2 + 4N}}{2(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i)}
\end{aligned}
$$

check the second derivative:

$$
\hat{\lambda} = -\frac{1}{2} + \frac{2N \pm \sqrt{(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i)^2 + 4N}}{2(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i)} < -\frac{1}{2}
$$

$$
\frac{\partial l(\lambda)}{\partial \lambda^2} = -\frac{N}{\lambda^2} + \frac{N}{(1-\lambda)^2} < 0
$$

Therefore, maximum likelihood estimator of the parameter λ is $\hat{\lambda}$

2 Problem 2

a)
The linear regression model $y = \mathbf{x}^\intercal \beta + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$
\begin{aligned}
L(\beta) = \max_\beta \prod_{i=1}^{N} Pr(y = y_i | x = x_i; \beta) = \max_\beta \prod_{i=1}^{N} (y - \mathbf{x}^\intercal \beta) \\
= \max_\beta \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(y - \mathbf{x}^\intercal \beta)^2 / 2} \\
l(\beta) = \ln \frac{N}{\sqrt{2\pi}} - \frac{(y - \mathbf{x}^\intercal \beta)^2}{2} \\
= \ln \frac{N}{\sqrt{2\pi}} - \frac{\mathbf{y}^\intercal \mathbf{y} + \mathbf{x}^\intercal \mathbf{x} \beta^2 - 2\mathbf{x}^\intercal \mathbf{y} \beta}{2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta} &= \mathbf{x}^\intercal \mathbf{y} - \mathbf{x}^\intercal \mathbf{x}\beta = 0 \\
\hat{\beta} &= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbf{y} \\
\mathbb{E}(\hat{\beta}) &= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbb{E}(\mathbf{y}) = (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbb{E}(\mathbf{x}\beta + \epsilon) \\
&= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal (\mathbf{x}\beta + \mathbb{E}(\epsilon)) \\
&= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbf{x}\beta \qquad [Ps : \mathbb{E}(\epsilon) = 0] \\
&= \beta \\
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbf{y}) = (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathrm{Var}(\mathbf{y})\mathbf{x}(\mathbf{x}^\intercal \mathbf{x})^{-1} \\
&= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x}^\intercal \mathbf{x}(\mathbf{x}^\intercal \mathbf{x})^{-1}\mathrm{Var}(\mathbf{y}) \\
&= (\mathbf{x}^\intercal \mathbf{x})^{-1}\mathrm{Var}(\mathbf{y}) \qquad [Ps : \mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\epsilon), \mathbf{x}, \beta \text{ are constant}] \\
&= (\mathbf{x}^\intercal \mathbf{x})^{-1}\sigma^2
\end{aligned}
$$

b)
If we know the full distribution of $\hat{\beta}$ with known mean and covariance matrix from the sample dataset, we can conduct Statistical hypothesis testing for β in the population. For example, we can do Significance Test for β under certain Significance level and calculate the P-Value of each entrance in β. Then we can get the test result for whether β is significant, so that we can reduce the data dimension and focus on the variables which bring more valuable information.

c)
We know $\hat{\beta}$ is normally distributed. and $\hat{y_*}$ is a linear function made of $\hat{\beta}$ and x_*, therefore, $\hat{y_*}$ is also normally distributed.

$$
\begin{aligned}
\mathbb{E}(\hat{y_*}) &= \mathbb{E}(\mathbf{x_*}^\intercal \hat{\beta}) = \mathbf{x_*}^\intercal\, \mathbb{E}(\hat{\beta}) = \mathbf{x_*}^\intercal \beta \\
\mathrm{Var}(\hat{y_*}) &= \mathrm{Var}(\mathbf{x_*}^\intercal \hat{\beta}) = \mathbf{x_*}\mathrm{Var}\hat{\beta}\mathbf{x_*}^\intercal = \mathbf{x_*}(\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x_*}^\intercal \sigma^2
\end{aligned}
$$

$$\text{Hence, } \hat{y_*} \sim \mathcal{N}(\mathbf{x_*}^\intercal \beta, \mathbf{x_*}(\mathbf{x}^\intercal \mathbf{x})^{-1}\mathbf{x_*}^\intercal \sigma^2)$$

d)
Knowing the distribution of the predictions of a model may come in handy when we want to evaluate new predictions that we make. Let us take predicting student performance in a math class as an example and assume that $\hat{y_*}$ is normally distributed. We might be interested to offer additional assistance to students who are struggling and more advanced materials for students who are doing well. Know- ing the distribution of our predictions can help us do that since we can compare how a prediction fairs with regard to the overall distribution. Furthermore, if this particular model were to be used as input for another model the information we gain could be applied as a weight. Thus, knowing the distribution of $\hat{y_*}$ can be beneficial to put new predictions into better context.