Solution 1:

a) Calculation of Pearson correlation coefficient of x_1 and x_2

$$\rho(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)(x_2^{(i)} - \overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_2^{(i)} - \overline{x}_2)^2}}$$

given the dataset:

	1	2	3	4	5	6	7	8	9	$\sum_{i=1}^{n}$
y	-7.79	-5.37	-4.08	-1.97	0.02	2.05	1.93	2.16	2.13	-10.92
x_1	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	0
x_2	0.95	0.57	0.29	-0.03	0.02	0.08	0.23	0.54	0.98	3.63

The individual differences to the means are:

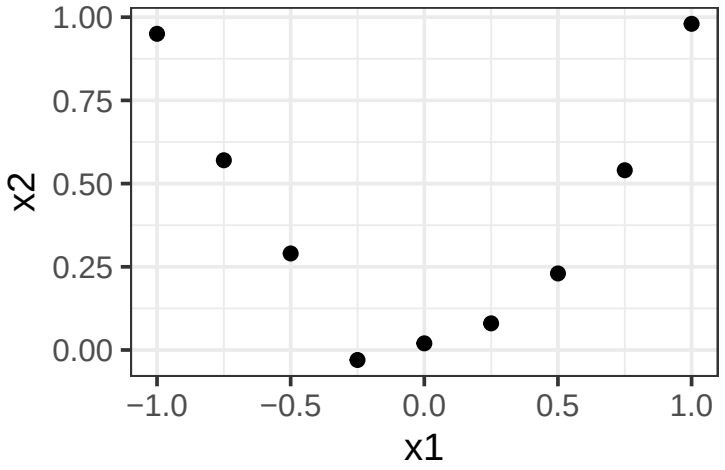
	1	2	3	4	5	6	7	8	9
$x_1^{(i)} - \overline{x}_1$	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
$x_2^{(i)} - \overline{x}_2$	0.55	0.17	-0.11	-0.43	-0.38	-0.32	-0.17	0.14	0.58

Then:

$$\begin{aligned}\rho(x_1, x_2) &= \frac{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)(x_2^{(i)} - \overline{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_1^{(i)} - \overline{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_2^{(i)} - \overline{x}_2)^2}} \\ &= \frac{-0.574 + -0.125 + 0.057 + 0.108 + 0 + -0.081 + -0.087 + 0.103 + 0.577}{2.086} = \frac{0.05}{2.086} = 0.002\end{aligned}$$

The Pearson correlation coefficient is close to 0. $\Rightarrow$ There is **no linear** relationship between x_1 and x_2, and they are not linearly correlated.

b) The scatter plot reveals that there is a strong non-linear/quadratic relationship between x_1 and x_2. The Pearson correlation coefficient is not suitable for detecting non-linear relationships.



c) The **mutual information (MI)** is more suitable for this data distribution. The formula for the mutual information is:

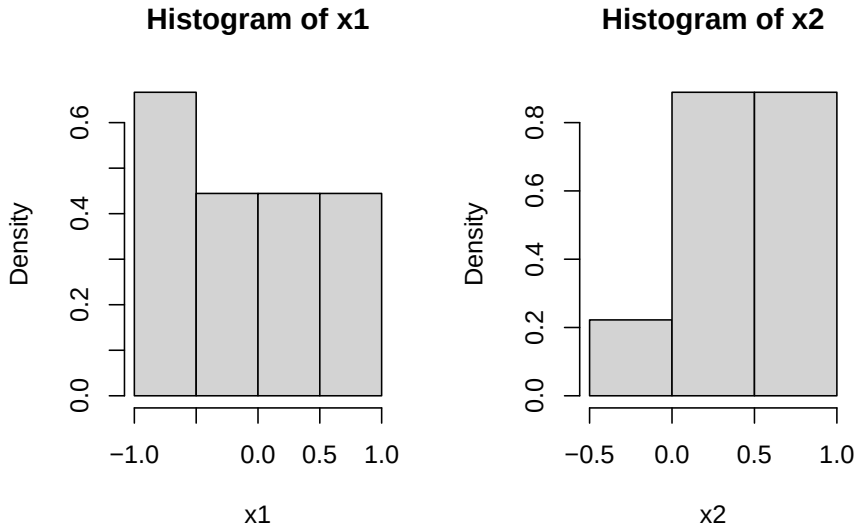
$$MI(x_1; x_2) = \mathbb{E}_{p(x_1,x_2)} \left[\log \left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right) \right] = \sum_{x_1} \sum_{x_2} p(x_1, x_2) \log \left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right) \tag{1}$$

Problem: We need a distribution, because the MI is defined using an expected value. This expected value can only be evaluated directly if we explicitly know the distribution of the data (which we do not in this case). If we simply use the empirical form (the sum in formula (1)) and plug in the 9 data points from above, we will obtain a very high MI, because we have too few samples. It would erroneously indicate an extremely strong dependency, but is not very informative in this case.

We therefore first have to estimate the whole distribution from our data points, and can then calculate the MI for this estimated distribution, which is an approximation of the true MI.
This approximation becomes better and better, the more data points one samples.

Solution 2* (Bonus Exercise):

One possible solution: Estimate the distribution using histograms with Gaussian kernel. This practically means: We split the ranges of the features into intervals (buckets) and calculate the empirical distributions w.r.t. these buckets:



Now, we take the mean values as replacement for the values in x_1 and x_2:

	1	2	3	4	5	6	7	8	9
$x_1^\star$	-0.75	-0.75	-0.75	-0.25	-0.25	0.25	0.25	0.75	0.75
$x_2^\star$	0.75	0.75	0.25	-0.25	0.25	0.25	0.25	0.75	0.75

Equivalently, one could also introduce categories for the single intervals and make x_1 and x_2 categorical variables. We then compute the table with joint and marginal distribution:

	$x_2^\star$	-0.25	0.25	0.75	p_{x_1}
$x_1^\star$	-0.75	0.00	0.11	0.22	0.33
	-0.25	0.11	0.11	0.00	0.22
	0.25	0.00	0.22	0.00	0.22
	0.75	0.00	0.00	0.22	0.22
	p_{x_2}	0.11	0.44	0.44	1.00

Now we can calculate the MI for this approximate distribution, which is very close to the MI of our original distribution:

$$\begin{aligned}
MI(x_1^\star; x_2^\star) &= \sum_{x_1^\star} \sum_{x_2^\star} p(x_1^\star, x_2^\star) \log\left(\frac{p(x_1^\star, x_2^\star)}{p(x_1^\star)p(x_2^\star)}\right) \\
&= 0 \log\left(\frac{0}{0.33 \cdot 0.11}\right) + 0.11 \log\left(\frac{0.11}{0.33 \cdot 0.44}\right) + 0.22 \log\left(\frac{0.22}{0.33 \cdot 0.44}\right) \\
&\quad + 0.11 \log\left(\frac{0.11}{0.22 \cdot 0.11}\right) + 0.11 \log\left(\frac{0.11}{0.22 \cdot 0.44}\right) + 0 \log\left(\frac{0}{0.22 \cdot 0.44}\right) \\
&\quad + 0 \log\left(\frac{0}{0.22 \cdot 0.11}\right) + 0.22 \log\left(\frac{0.22}{0.22 \cdot 0.44}\right) + 0 \log\left(\frac{0}{0.22 \cdot 0.44}\right) \\
&\quad + 0 \log\left(\frac{0}{0.22 \cdot 0.11}\right) + 0 \log\left(\frac{0}{0.22 \cdot 0.44}\right) + 0.22 \log\left(\frac{0.22}{0.22 \cdot 0.44}\right) \\
&\approx 0.603
\end{aligned}$$

$\Rightarrow$ The MI of around 0.6 shows that there is a clear dependency.

Solution 3:

Prerequisites:

We will need a few general facts about least-squares linear regression for this exercise:

1. The formulae for estimating the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of the linear regression model. These formulae can be calculated from the model equation $\hat{f}_{LM}(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and the loss equation $SSE_{LM} = \sum_{i=1}^n \left(y^{(i)} - \hat{f}_{LM}(x^{(i)})\right)^2$ using standard calculus.

$$\begin{aligned}
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \iff \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 = \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}
\end{aligned}$$

   One can prove (with standard multivariate calculus) that these values of the parameters are the unique minimizers of the loss function. It follows that:

$$(\hat{y}^{(i)} - \bar{y}) = \hat{\beta}_0 + \hat{\beta}_1 x^{(i)} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \hat{\beta}_1 (\hat{x}^{(i)} - \bar{x}).$$

2. With the notation

$$\begin{aligned}
SSE_{LM} &= \sum_{i=1}^n (y^{(i)} - \hat{f}_{LM}(x^{(i)}))^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 && \text{for the sum of squared errors (the loss) from the regression,} \\
SSE_c &= \sum_{i=1}^n (y^{(i)} - \bar{y})^2 && \text{for the total sum of squares / the variance of the data points,} \\
SSE_{LM-c} &= \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 = \sum_{i=1}^n (\hat{f}_{LM}(x^{(i)}) - \bar{y})^2 && \text{for the total sum of squares / the variance of the predictions,}
\end{aligned}$$

   the following holds:

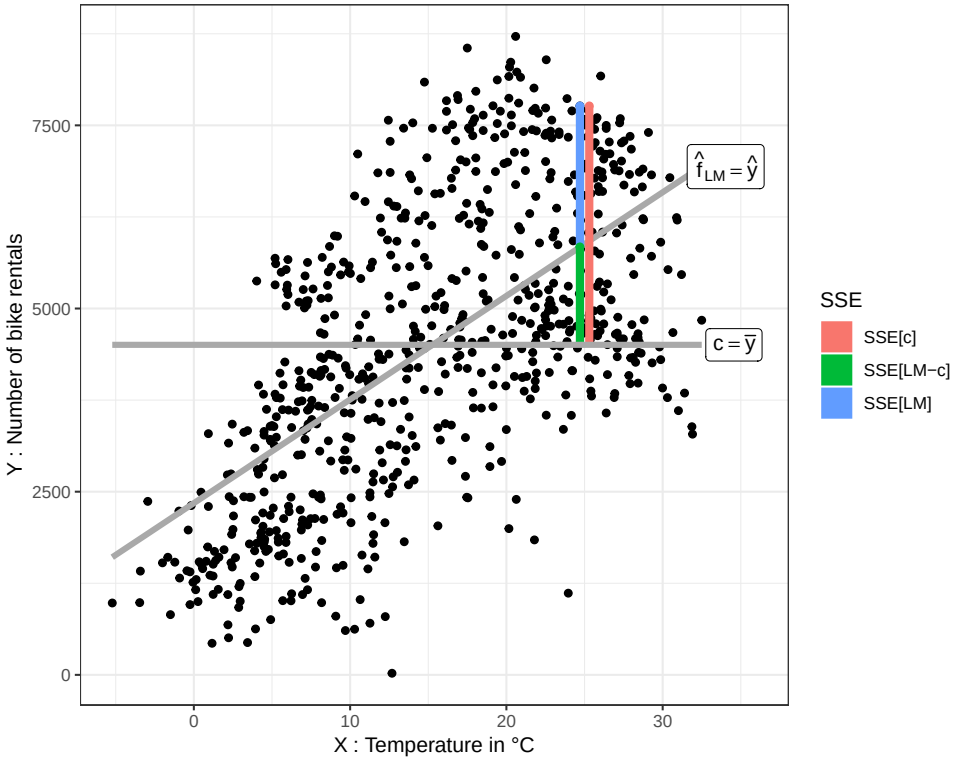
$$SSE_c = SSE_{LM} + SSE_{LM-c} \iff \sum_{i=1}^n (y^{(i)} - \bar{y})^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 \tag{2}$$

   This basically means that the model predictions $\hat{y}_i$ and the residuals $(y_i - \hat{y}_i)$ are uncorrelated. (The two vectors are "orthogonal" to each other.) A proof for formula (2) is given below.

3. We can use (2) to rewrite the coefficient of determination as follows:

$$R^2 = 1 - \frac{SSE_{LM}}{SSE_c} = \frac{SSE_{LM-c}}{SSE_c} = \frac{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}$$

An illustration of this relation:



Proof of equation (2)**:**

We want to show:

$$\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2 = \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2.$$

Because of

$$\begin{aligned}\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2 &= \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)} + \hat{y}^{(i)} - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + (\hat{y}^{(i)} - \bar{y})^2 + 2(y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y}) \\ &= \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 + \sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{y})^2 + 2\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y}),\end{aligned}$$

it is sufficient to show that the following covariance vanishes:

$$\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y}) = 0$$

This is indeed the sample covariance between $(y^{(i)} - \hat{y}^{(i)})$ and $\hat{y}^{(i)}$ (or equivalently between $(y^{(i)} - \hat{y}^{(i)})$ and $(\hat{y}^{(i)} - \bar{y})$), because $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}^{(i)} = \bar{y}$ is the average of $\hat{y}^{(i)}$ and therefore $(y^{(i)} - \hat{y}^{(i)})$ has an average of 0.

We start with the formula for $\hat{\beta}_1$:

$$\begin{aligned}
& & \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2} \\
\Rightarrow \quad & \hat{\beta}_1 \sum_{i=1}^n (x^{(i)} - \bar{x})^2 &= \sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y}) \\
\Leftrightarrow \quad & 0 &= \sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x^{(i)} - \bar{x})^2 = \sum_{i=1}^n \left((x^{(i)} - \bar{x})(y^{(i)} - \bar{y}) - \hat{\beta}_1 (x^{(i)} - \bar{x})^2 \right) \\
& &= \sum_{i=1}^n \left((y^{(i)} - \bar{y}) - \hat{\beta}_1 (x^{(i)} - \bar{x}) \right) (x^{(i)} - \bar{x}) = \sum_{i=1}^n \left((y^{(i)} - \bar{y}) - (\hat{y}^{(i)} - \bar{y}) \right) (x^{(i)} - \bar{x}) \\
& &= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(x^{(i)} - \bar{x}) \\
\Rightarrow \quad & 0 &= \hat{\beta}_1 \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(x^{(i)} - \bar{x}) = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(\hat{\beta}_1 (x^{(i)} - \bar{x})) \\
& &= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})(\hat{y}^{(i)} - \bar{y})
\end{aligned}$$

(proof of (2)) □

Proof of $R^2 = \rho^2$:

We start with R^2 and plug in the functional equation from the regression for the data points and for the sample averages. A little more calculation yields the desired result:

$$\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{\beta}_1 (x^{(i)} - \bar{x}))^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} \\
&= \left(\frac{\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2} \right)^2 \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} = \frac{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y}) \right)^2}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2 \sum_{i=1}^n (x^{(i)} - \bar{x})^2} \frac{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}{\sum_{i=1}^n (y^{(i)} - \bar{y})^2} \\
&= \frac{\left(\sum_{i=1}^n (x^{(i)} - \bar{x})(y^{(i)} - \bar{y}) \right)^2}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2 \sum_{i=1}^n (y^{(i)} - \bar{y})^2} \\
&= \rho^2
\end{aligned}$$

This shows $R^2 = \rho^2$, which completes the proof. □

Note that this result is only valid for simple linear regression, only for the case of one independent variable, and also only for regression with ordinary least squares (OLS). For multiple regression, the coefficient of determination is defined differently and does not necessarily equal the square of the Pearson correlation coefficient.

Similar proofs together with more information:
https://statproofbook.github.io/P/slr-rsq.html
https://math.stackexchange.com/questions/129909/correlation-coefficient-and-determination-coefficient

Solution 4:

First step: We need to calculate the respective second order partial derivative.

Problem: The function $f(\mathbf{x}) = 2x_1 + 3x_2 - x_1|x_2|$ is not differentiable for $x_2 = 0$. Hence, different cases need to be considered:

$$\text{Case 1: } x_2 > 0 \quad ; \quad \text{Case 2: } x_2 < 0 \quad ; \quad \text{Case 3: } x_2 = 0$$

Case 1: $x_2 > 0$

$$\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ (2x_1 + 3x_2 - x_1 x_2)\right)^2 = \left(\frac{\partial}{\partial x_2}\ (2 - x_2)\right)^2 = (-1)^2 = 1 > 0$$

Case 2: $x_2 < 0$

$$\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ (2x_1 + 3x_2 - x_1(-x_2))\right)^2 = \left(\frac{\partial}{\partial x_2}\ (2 + x_2)\right)^2 = 1^2 = 1 > 0$$

Case 3: $x_2 = 0$

Whenever $x_1 \neq 0$, we know that the function has the form $f(\mathbf{x}) = 2x_1 + 3x_2 + x_1 x_2$ left from the point $(x_1, 0)$, and the form $f(\mathbf{x}) = 2x_1 + 3x_2 - x_1 x_2$ right from the point, hence at $(x_1, 0)$ it is not differentiable with respect to x_2, and therefore our definition cannot be applied.

However, for the point $(x_1, x_2) = 0$, the function takes the form $f(\mathbf{x}) = 2x_1$ on both sides in x_1 direction, and the form $f(\mathbf{x}) = 3x_2$ on both sides in x_2 direction. We can conclude that f is twice differentiable at $(0, 0)$ with $\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} = 0$ at this point.

Second step: Computing the expected value.

The squared second derivative is a constant 1 or 0, if defined. Hence, as long as we have $\mathbb{P}(x_2 \neq 0) > 0$, we will always have that the expected value is > 0, and therefore, x_1 and x_2 interact with each other.

On the other hand, in the case where $\mathbb{P}(x_2 \neq 0) = 0$, it follows that $\mathbb{P}(x_2 = 0) = 1$, which means that x_2 is the constant 0 almost surely (with 100% probability). In this case (when one variable is constant with probability 1), it does not make much sense to analyze for interactions.

$\Rightarrow$ In this example, there is an interaction between x_1 and x_2.

Solution 5* (Bonus Exercise):

a) As in exercise 4, we first calculate the second derivative. Again, we have to distinguish three cases with respect to x_1.

Case 1: $x_1 > 0$

$$\left(\frac{\partial^2 g(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ \left(0.01e^{x_1^2} + \sin(x_2)\sqrt{x_1} - 1.5x_2^3\right)\right)^2 = \left(\frac{\partial}{\partial x_2}\ \left(0.01e^{x_1^2} \cdot 2x_1 + \sin(x_2)\frac{1}{2\sqrt{x_1}}\right)\right)^2$$
$$= \left(\frac{\cos(x_2)}{2\sqrt{x_1}}\right)^2 = \frac{\cos^2(x_2)}{2x_1} \geq 0,$$

because we assumed $x_1 > 0$ in this case.

Case 2: $x_1 < 0$

$$\left(\frac{\partial^2 g(\mathbf{x})}{\partial x_1 \partial x_2}\right)^2 = \left(\frac{\partial^2}{\partial x_1 \partial x_2}\ \left(0.01e^{x_1^2} + \sin(x_2)\sqrt{0} - 1.5x_2^3\right)\right)^2 = \left(\frac{\partial}{\partial x_2}\ \left(0.01e^{x_1^2} \cdot 2x_1\right)\right)^2 = 0$$

Case 3: $x_1 = 0$

In this case, for any point $(0, x_2)$, the right-side limit of the partial derivative in x_1-direction is ∞, i.e. the partial derivative has a singularity in this point, whereas the left-side limit is 0. Therefore, g can never be partially differentiable in x_1 direction in such a point (only differentiable from the left-hand side), and therefore the required second-order derivative also does not exist.

Expected value:

Very similarly to exercise 4, we can conclude again that the expected value will be greater than 0 as long as $\mathbb{P}(X_1 > 0) > 0$, i.e. as long as there is a non-zero chance of x_1 being greater than 0.
On the other hand, for any distribution with $\mathbb{P}(X_1 \geq 0) = 0$, that is with $\mathbb{P}(X_1 < 0) = 1$ (meaning that X_1 is

negative with a probability of 1), we actually get an expected value of 0, because in any region with probability > 0, the second derivative is 0, and the expected value of 0 is 0.

Examples for two probability distributions:

For this example we will use an exponential probability distribution with an arbitrary parameter $\lambda > 0$. We know that if some random variable X is distributed exponentially with parameter λ (written $X \sim \text{Exp}_\lambda$), then $\mathbb{P}(X > 0) = 1$ and $P(X < 0) = 0$. You can also construct this example with any other distribution which is concentrated on the positive half-axis, e.g. a geometric distribution or a binomial distribution.

With interactions: Choose $X_1 \perp\!\!\!\perp X_2$ (this means that X_1 and X_2 are independent), $X_1 \sim \text{Exp}_\lambda$ exponentially and $X_2 \sim \mathcal{N}(\mu, \sigma)$ normally distributed. (You can also choose any other probability distribution for X_2.) Then we get from the independence that $\mathbb{P}(X_1 > 0) = 1$, and therefore

$$\mathbb{E}_{\mathbf{X}}\left[\left(\frac{\partial^2 g(\mathbf{X})}{\partial x_i \partial x_j}\right)^2\right] = \mathbb{E}_{\mathbf{X}}\left[\frac{\cos^2(X_2)}{2X_1}\right] > 0,$$

because the function inside is greater than 0 on the half-space $\{x_1 > 0\}$, on which the distribution is concentrated. So there are interactions in this case.

Without interactions: Again choose $X_1 \perp\!\!\!\perp X_2$ and $X_2 \sim \mathcal{N}(\mu, \sigma)$ normally distributed, but now choose $-X_1 \sim \text{Exp}_\lambda$, so X_1 follows the symmetric mirror of an exponential distribution. In this case we have $\mathbb{P}(X_1 < 0) = 1$. Hence

$$\mathbb{E}_{\mathbf{X}}\left[\left(\frac{\partial^2 g(\mathbf{X})}{\partial x_i \partial x_j}\right)^2\right] = \mathbb{E}_{\mathbf{X}}\left[0\right] = 0,$$

because $\mathbb{P}(X_1 > 0) = 0$. Therefore, no interactions are present for this distribution.

Summary:

In a nutshell, the function g exhibits interactions only on the half-space $\{(x_1, x_2) | x_1 > 0\}$ of $\mathbb{R}^2$. Therefore, there are interactions present if and only if the data distribution is at least partially contained in this half-space, so there is at least some data inside this half-space.

b) For both distributions in part a), the function f from exercise 4 would exhibit interactions, since its second partial derivative squared is 1 whenever $x_2 \neq 0$, so we would take the expected value of 1, which is 1. This also holds when swapping the roles of x_1 and x_2, so that x_2 is exponentially distributed and concentrated on one half-space.

We can even prove that no other distribution could change this result. As discussed in the solution for exercise 4, the function f always contains interactions except if $\mathbb{P}(X_2 = 0) = 1$, meaning that no other value of x_2 has positive probability. Hence, the only kind of distribution where f exhibits no interactions is that where all probability of x_2 is concentrated on the single point 0, and there is no distribution continuous in both x_1 and x_2 where f would contain no interactions.