Note: I assume the 'Amount' in both datasets to be spending amount in this exercise.

Challenge 1:

license-plate, TransactionID and Amount columns are not the common fields between 'loyalty' and 'transactions', so these 3 fields are not taken into consideration as mapping keys.

1. Duplicates are found in name, city, phone-number and email columns, so these 4 fields are not the best joining key.
2. Both datasets have id column, and there is no duplicate in these columns in both 'loyalty' and 'transactions', we can consider id as the primary key. Hence, the joining key is id.
3. No null values found in both id columns.
4. The datasets are joined using 'inner join' and the newly created table contains all 10000 id(s) from both 'loyalty' and 'transactions', meaning to say both datasets have the same ids.

Business use case:

Assuming that 'Amount' is the customer's spending amount, we can classify them into different groups by the spending amount. So we now have the high level spending behaviour of the customers, which are high spending, moderate spending and low spending classes. We can sell different products (with different price range) to each group of them depending on their spending behaviour. Since we have their PII data, such as emails and phone numbers, we can blast communications via SMS or eDM to target them for different products.

Besides, we can derive additional info from the existing data.

a) Derive **email type** from email address. The unique email types in the datasets are:

verizon.com
outlook.com
mail.com
zoho.com
yahoo.com
xfinity.com
protonmail.com
hotmail.com
yandex.com
comcast.net
att.com
aol.com
gmail.com

b) Derive **area code** from phone-number. The first 3 digits of the phone-number are the area code.
(Note: Phone numbers in the United States typically consist of 11 digits — the 1-digit country code, a 3-digit area code and a 7-digit telephone number. The area codes represent the geographic region to which a phone number belongs)

Area codes represent the regional location of a phone number, your business phone number is a direct indicator of where you are located. The location of your business contributes to how your customers get in touch with you, or even perceive you. With a local phone number, your regional customer base can contact you without incurring extra charges and vice versa. If we look at it from another way round, we can target the customers by area to save cost.

From the additional info derived, we can know for each spending class, which email type they prefer and where do they reside. For example, the lowest spending class (spending amount: 0 - 250) use verizon.com and comcast.net the most.

binned_amount	1. 0-250

Row Labels	Count of id	Count of id2
verizon.com	113	9.1%
comcast.net	107	8.6%
yandex.com	105	8.5%
hotmail.com	99	8.0%
zoho.com	94	7.6%
aol.com	94	7.6%
xfinity.com	92	7.4%
att.com	92	7.4%
outlook.com	91	7.4%
yahoo.com	90	7.3%
gmail.com	88	7.1%
mail.com	87	7.0%
protonmail.com	86	6.9%
Grand Total	**1238**	**100.0%**

While the highest spending class (spending amount: >1750 – 2000) use aol.com and mail.com the most.

binned_amount	8. >1750-2000

Row Labels	Count of id	Count of id2
aol.com	106	8.6%
mail.com	103	8.4%
xfinity.com	101	8.2%
comcast.net	101	8.2%
verizon.com	99	8.1%
hotmail.com	98	8.0%
zoho.com	97	7.9%
outlook.com	95	7.7%
yandex.com	92	7.5%
protonmail.com	86	7.0%

yahoo.com	85	6.9%
gmail.com	84	6.8%
att.com	82	6.7%
Grand Total	**1229**	**100.0%**

If we do not know the spending amount of a customer but we manage to get the PII data (email address/phone number), we can still do extrapolation from there to expand the targeting base.

Challenge 2 :

Please refer to the power point slides.

Slide 1:

Out of a total of 10K customers, the area codes of the top 5 areas that they reside in are 591, 463, 396, 504 and 727.

The most used emails are Hotmail, Verizon, Zoho, Comcast and Yandex.

The spending amount among the customers are evenly distributed in general, there are around 1.2K customers in each spending class.

Slide 2:

Deep diving into the lowest spending classs, we can see that most of them reside in 728 area and the most preferred email by this group is Verizon.

Looking at the highest spending class, they mostly reside in 828, 531 and 442 and their most used email is AOL.

Challenge 3:

Data enrichment:

Enriching the existing data that we have with external data, meaning to say based on whichever data we have, we can further expand the data by gaining more info from external sources which can bring more values or insights.

Use case:

Since now we have area codes, we can further enrich the data by scraping the region and time zone for each area code. From here, we can further analyse the customer groups based on region and to see if there are any similarities among the customers within the same region/time zone.

Beside analysis purpose, we can use the scaped data for targeting purpose. We can estimate and choose the best time to call and sell to a customer based on the time zone that the customer resides.

For web scraping script, please refer to https://github.com/kheng2yee/takehometest