1. I find **<u>the product ID can repeat</u>** – we can find the most popular product and what's the characteristic of the customers who bought it.
   P00025442 is the most popular product which has the highest sum purchase number; while P0086242 has the highest average purchase number
   So, this two product can be our target product to do further discussion.

   1) What's the characteristic of the customers who bought it.

2. Change columns into factors by lapply() should not include NA, or it will cause mistakes.

3. Describe data:

   (1) Male cost much than female

Gender <fctr>	sum(sumP/1) <dbl>	mean(sumP/1) <dbl>
F	1164624021	699054.0
M	3853044357	911963.2

2 rows

   (2) The male whose age is **<u>among 26-35</u>** bought the most

Gender <fctr>	Age <fctr>	sum(sumP) <int>	mean(sumP) <dbl>
F	0-17	41826615	536238.7
F	18-25	202209450	704562.5
F	26-35	433857680	796069.1
F	36-45	239010480	717749.2
F	46-50	114796993	630752.7
F	51-55	87972407	619524.0
F	55+	44950396	454044.4
M	0-17	90832391	648802.8
M	18-25	699459830	894449.9
M	26-35	1565891426	1038389.5

Gender <fctr>	Age <fctr>	sum(sumP) <int>	mean(sumP) <dbl>
M	36-45	771639085	925226.7
M	46-50	298621230	855648.2
M	51-55	273935949	808070.6
M	55+	152664446	559210.4

(3) City_category

City_Category <fctr>	sum(sumP) <int>	mean(sumP) <dbl>
A	1295668797	1239874.4
B	2083431612	1220522.3
C	1638567969	522003.2

(4) Stay _in_city_years

Stay_In_Current_City_Years <fctr>	sum(sumP) <int>	mean(sumP) <dbl>
0	672505429	871121.0
1	1763243917	845275.1
2	934676626	816311.5
3	872531130	891247.3
4+	774711276	852267.6

(5) Marital_status + gender

Gender <fctr>	Marital_Status <fctr>	sum(sumP/1) <dbl>	mean(sumP) <dbl>
F	0	673815717	711526.6
F	1	490808304	682626.3
M	0	2292473783	928127.0
M	1	1560570574	889214.0