
1차 데이터 변형

```
# 불필요한 컬럼 제거 ('Payment_sequential', 'Product_weight_g',  
'Product_length_cm', 'Product_height_cm', 'Product_width_cm')  
tmp1 = temp.groupby(['Review_id', 'Order_id', 'Review_score',  
'Review_creation_date', 'Review_answer_timestamp', 'Order_item_id',  
'Product_id', 'Seller_id', 'Price', 'Freight_value', 'Customer_id',  
'Order_status', 'Order_purchase_timestamp', 'Order_delivered_carrier_date',  
'Order_delivered_customer_date', 'Order_estimated_delivery_date',  
'Payment_type', 'Payment_installments', 'Product_category_name',  
'Customer_unique_id', 'Customer_zipcode_prefix', 'Customer_city',  
'Customer_state', 'Seller_zipcode_prefix', 'Seller_city',  
'Seller_state'], as_index=False)[['Payment_type', 'Payment_value']].sum()
```

tmp1

```

tmp1 = temp.groupby(['Review_id', 'Order_id', 'Review_score', 'Review_creation_date',
                    'Review_answer_timestamp', 'Order_item_id', 'Product_id', 'Seller_id',
                    'Price', 'Freight_value', 'Customer_id', 'Order_status',
                    'Order_purchase_timestamp', 'Order_delivered_carrier_date',
                    'Order_delivered_customer_date', 'Order_estimated_delivery_date',
                    'Payment_type', 'Payment_installments',
                    'Product_category_name', 'Customer_unique_id', 'Customer_zipcode_prefix', 'Customer_city',
                    'Customer_state', 'Seller_zipcode_prefix', 'Seller_city',
                    'Seller_state'],as_index=False)[['Payment_type', 'Payment_value']].sum()

```

tmp1

✓ 0.8s

	Review_id	Order_id	Review_score	Review_creation_date	Review_answer_timestamp	Order_item_id	
0	REVIEW_00000	ORDER_01674	4	2019-01-18 00:00:00	2019-01-18 21:46:59	1	F
1	REVIEW_00000	ORDER_01674	4	2019-01-18 00:00:00	2019-01-18 21:46:59	2	F
2	REVIEW_00001	ORDER_80140	5	2019-03-10 00:00:00	2019-03-11 03:05:13	1	F
3	REVIEW_00002	ORDER_69816	5	2019-02-17 00:00:00	2019-02-18 14:36:24	1	F
4	REVIEW_00003	ORDER_24398	5	2018-04-21 00:00:00	2018-04-21 22:02:06	1	F
...
102834	REVIEW_87298	ORDER_58840	5	2019-07-07 00:00:00	2019-07-14 17:18:30	1	F
102835	REVIEW_87299	ORDER_75162	5	2018-12-09 00:00:00	2018-12-11 20:06:42	1	F
102836	REVIEW_87300	ORDER_08690	5	2019-03-22 00:00:00	2019-03-23 09:10:43	1	F
102837	REVIEW_87301	ORDER_25681	4	2019-07-01 00:00:00	2019-07-02 12:59:13	1	F
102838	REVIEW_87302	ORDER_45326	1	2018-07-03 00:00:00	2018-07-03 21:01:49	1	F

102839 rows × 27 columns

```
tmp1.shape
```

```
✓ 0.0s
```

```
(102839, 27)
```

```
len(tmp1[tmp1.Order_item_id > 1])
```

```
✓ 0.0s
```

```
12801
```

```
len(tmp1[tmp1.Order_id == 'ORDER_68748'])
```

```
✓ 0.0s
```

```
2
```

2차 데이터 변형

```
tmp2 = tmp1.groupby(['Review_id', 'Order_id', 'Review_score',  
    'Review_creation_date', 'Review_answer_timestamp', 'Product_id', 'Seller_id',  
    'Price', 'Freight_value', 'Customer_id', 'Order_status',  
    'Order_purchase_timestamp', 'Order_delivered_carrier_date',  
    'Order_delivered_customer_date', 'Order_estimated_delivery_date',  
    'Payment_type', 'Payment_installments', 'Product_category_name',  
    'Customer_unique_id', 'Customer_zipcode_prefix', 'Customer_city',  
    'Customer_state', 'Seller_zipcode_prefix', 'Seller_city',  
    'Seller_state'], as_index=False).Order_item_id.count()
```

```
tmp2 = tmp2.rename(columns={"Order_item_id" : "order_count"})
```

```
tmp2
```

```

tmp2 = tmp1.groupby(['Review_id', 'Order_id', 'Review_score', 'Review_creation_date', 'Review_answer_timestamp', 'Product_id', 'Seller_id', 'Price', 'Freight_value',
                    'Customer_id', 'Order_status', 'Order_purchase_timestamp', 'Order_delivered_carrier_date', 'Order_delivered_customer_date',
                    'Order_estimated_delivery_date', 'Payment_type', 'Payment_installments', 'Product_category_name', 'Customer_unique_id', 'Customer_zipcode_prefix',
                    'Customer_city', 'Customer_state', 'Seller_zipcode_prefix', 'Seller_city', 'Seller_state'], as_index=False).Order_item_id.count()

tmp2 = tmp2.rename(columns={"Order_item_id" : "order_count"})

tmp2

```

✓ 0.7s

	Review_id	Order_id	Review_score	Review_creation_date	Review_answer_timestamp	Product_id	Seller_id	Price	Freight_value	Customer_id	Order_status	Order
0	REVIEW_00000	ORDER_01674	4	2019-01-18 00:00:00	2019-01-18 21:46:59	PRODUCT_21853	SELLER_0286	185.00	13.63	CUSTOMER_38995	delivered	
1	REVIEW_00001	ORDER_80140	5	2019-03-10 00:00:00	2019-03-11 03:05:13	PRODUCT_18124	SELLER_0262	79.79	8.30	CUSTOMER_81808	delivered	
2	REVIEW_00002	ORDER_69816	5	2019-02-17 00:00:00	2019-02-18 14:36:24	PRODUCT_07372	SELLER_2445	149.00	45.12	CUSTOMER_27108	delivered	
3	REVIEW_00003	ORDER_24398	5	2018-04-21 00:00:00	2018-04-21 22:02:06	PRODUCT_22159	SELLER_2445	179.99	42.85	CUSTOMER_62103	delivered	
4	REVIEW_00004	ORDER_70366	5	2019-03-01 00:00:00	2019-03-02 10:26:53	PRODUCT_19699	SELLER_1555	1199.00	134.25	CUSTOMER_57462	delivered	
...
93476	REVIEW_87298	ORDER_58840	5	2019-07-07 00:00:00	2019-07-14 17:18:30	PRODUCT_25263	SELLER_0766	226.77	61.20	CUSTOMER_06636	delivered	
93477	REVIEW_87299	ORDER_75162	5	2018-12-09 00:00:00	2018-12-11 20:06:42	PRODUCT_03024	SELLER_2634	199.99	9.77	CUSTOMER_48218	delivered	
93478	REVIEW_87300	ORDER_08690	5	2019-03-22 00:00:00	2019-03-23 09:10:43	PRODUCT_24946	SELLER_1192	215.97	15.59	CUSTOMER_86112	delivered	
93479	REVIEW_87301	ORDER_25681	4	2019-07-01 00:00:00	2019-07-02 12:59:13	PRODUCT_20058	SELLER_0642	50.95	15.46	CUSTOMER_54637	delivered	
93480	REVIEW_87302	ORDER_45326	1	2018-07-03 00:00:00	2018-07-03 21:01:49	PRODUCT_01615	SELLER_1845	32.90	7.78	CUSTOMER_83783	delivered	

93481 rows x 26 columns

```

tmp2.shape

```

✓ 0.0s

(93481, 26)

중복 데이터 확인

#아래 캡처 순서대로

```
tmp2[tmp2.Review_id.duplicated()]
```

```
tmp2[tmp2.Review_id == 'REVIEW_00005']
```

```
order_items[order_items.Order_id == 'ORDER_23038']
```

```
tmp2[tmp2.Review_id.duplicated()]
```

✓ 0.0s

	Review_id	Order_id	Review_score	Review_creation_date	Review
6	REVIEW_00005	ORDER_23038	1	2019-04-13 00:00:00	
15	REVIEW_00013	ORDER_24124	3	2018-04-30 00:00:00	
32	REVIEW_00029	ORDER_56423	5	2018-07-19 00:00:00	
39	REVIEW_00035	ORDER_74871	1	2018-04-21 00:00:00	
52	REVIEW_00047	ORDER_37372	1	2019-01-28 00:00:00	
...
93408	REVIEW_87234	ORDER_28301	5	2019-06-09 00:00:00	
93420	REVIEW_87245	ORDER_10322	3	2019-03-09 00:00:00	
93444	REVIEW_87268	ORDER_24901	5	2019-06-20 00:00:00	
93455	REVIEW_87278	ORDER_65421	5	2019-06-30 00:00:00	
93467	REVIEW_87289	ORDER_19207	5	2018-03-08 00:00:00	

6309 rows × 26 columns

```
tmp2[tmp2.Review_id == 'REVIEW_00005']
```

✓ 0.0s

	Review_id	Order_id	Review_score	Review_creation_date	Review_answer_timestamp	Product_id	Seller_id	Price	Freight_value	CU
5	REVIEW_00005	ORDER_23038	1	2019-04-13 00:00:00	2019-04-16 00:39:37	PRODUCT_15917	SELLER_1678	99.9	13.2	CU
6	REVIEW_00005	ORDER_23038	1	2019-04-13 00:00:00	2019-04-16 00:39:37	PRODUCT_28359	SELLER_0042	119.0	4.4	CU

```
order_items[order_items.Order_id == 'ORDER_23038']
```

✓ 0.0s

	Order_id	Order_item_id	Product_id	Seller_id	Price	Freight_value
69704	ORDER_23038	1	PRODUCT_15917	SELLER_1678	99.9	13.2
69705	ORDER_23038	2	PRODUCT_15917	SELLER_1678	99.9	13.2
69706	ORDER_23038	3	PRODUCT_28359	SELLER_0042	119.0	4.4
69707	ORDER_23038	4	PRODUCT_15917	SELLER_1678	99.9	13.2