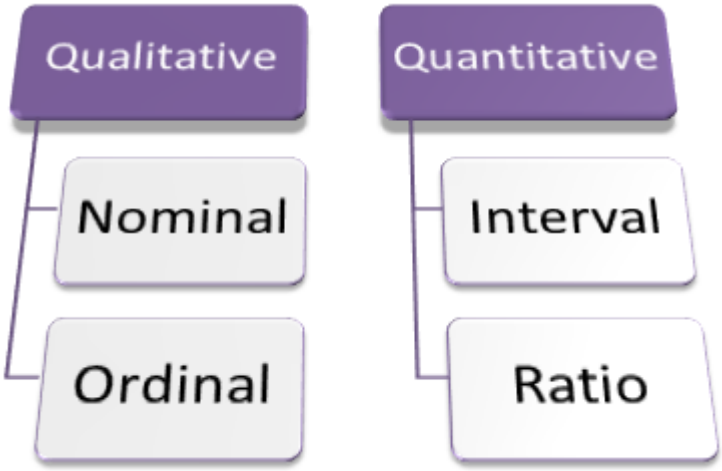


Data Measurement Type	Property	Central Tendency Measure	Allowed math	Example
NOMINAL	grouping	mode	-	color
ORDINAL	order	mode, median	monotone transformations	education level
INTERVAL	equal intervals	mode, median, mean	linear transformations	temperature in °C
RATIO	meaningful zero	mode, median, mean	scaling transformations	price

In [1]:
```
import pandas as pd
import numpy as np
train = pd.read_csv('train.gz')
test = pd.read_csv('test.gz')
```

In [2]:
```
# train test
fullset = pd.concat([train,test],ignore_index=True)
```

In [3]:
```
def meta(train,test,missing_values = -1,cols_ignore_missing = []):

    df = pd.concat([train,test]).reset_index(drop=True).fillna('未知')
    data = []
    for col in df.columns:
        # 定义role
        if col == 'target':
            role = 'target'
        elif col == 'id':
            role = 'id'
        else:
            role = 'feature'

        # 定义category
        if 'ind' in col:
            category = 'individual'
        elif 'car' in col:
            category = 'car'
        elif 'calc' in col:
            category = 'calculated'
        elif 'reg' in col:
            category = 'region'
        else:
            category = 'other'


        # 定义 level of measurements
        if 'bin' in col or col == 'target':
            level = 'binary'
        elif 'cat' in col[-3:] or col == 'id':
            level = 'nominal'
        elif df[col].dtype == 'float64' and df[col].replace(missing_values,
            level = 'interval'
        elif df[col].dtype == 'float64' and df[col].replace(missing_values,
            level = 'ratio'
        elif df[col].dtype == 'int64':
            level = 'ordinal'

        # 定义 data type
        dtype = df[col].dtype

        # 定义 unique
        if col == 'id' or df[col].dtype == 'float64':
            uniq = 'Ignore'
        else:
            if col in cols_ignore_missing:
                uniq = df[col].nunique()
            else:
                uniq = df[col].replace({missing_values:np.nan}).nunique()

        # 定义 cardinality
        if uniq == 'Ignore':
            cardinality = 'Ignore'
        elif uniq <= 10:
            cardinality = 'Low Cardinality'
        elif uniq <= 30:
```

```
            cardinality = 'Medium Cardinality'
        else:
            cardinality = 'High Cardinality'

        # 定义 missing
        if col in cols_ignore_missing:
            missing = 0
        else:
            missing = sum(df[col] == missing_values)

        # 定义 missing percent
        missing_percent = f'{missing}({round(missing*100/len(df),2)}%)'

        # 定义 imputation
        if missing > df.shape[0]*0.4:
            imputation = 'remove'
        elif missing > 0:
            if level == 'binary' or level == 'nominal':
                imputation = ('mode')
            if level == 'ordinal':
                imputation = ('mode','median')
            if level == 'interval' or level == 'ratio':
                imputation = ('mode','median','mean')
        else:
            imputation = "No Missing"

        # 定义 keep
        keep = True
        if col  == 'id' or imputation == 'remove':
            keep = False
        col_dict = {
            'colname': col,
            'role': role,
            'category': category,
            'level': level,
            'dtype': dtype,
            'cardinality': uniq,
            'cardinality_level':cardinality,
            'missing': missing,
            'missing_percent': missing_percent,
            'imputation':imputation,
            'keep': keep,
        }
        data.append(col_dict)
    meta = pd.DataFrame(data, columns=list(col_dict.keys()))
    meta.set_index('colname', inplace=True)

    return meta
```

In [4]:
```
metadata = meta(train,test)
```

In [5]:
```
metadata
```

Out[5]:

colname	role	category	level	dtype	cardinality	cardinality_level	missing	missing_p
id	id	other	nominal	int64	Ignore	Ignore	0	0
target	target	other	binary	object	3	Low Cardinality	0	0
ps_ind_01	feature	individual	ordinal	int64	8	Low Cardinality	0	0
ps_ind_02_cat	feature	individual	nominal	int64	4	Low Cardinality	523	523(0
ps_ind_03	feature	individual	ordinal	int64	12	Medium Cardinality	0	0
ps_ind_04_cat	feature	individual	nominal	int64	2	Low Cardinality	228	228(0
ps_ind_05_cat	feature	individual	nominal	int64	7	Low Cardinality	14519	14519(0
ps_ind_06_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_07_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_08_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_09_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_10_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_11_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_12_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_13_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_14	feature	individual	ordinal	int64	5	Low Cardinality	0	0
ps_ind_15	feature	individual	ordinal	int64	14	Medium Cardinality	0	0
ps_ind_16_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_17_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_ind_18_bin	feature	individual	binary	int64	2	Low Cardinality	0	0
ps_reg_01	feature	region	ratio	float64	Ignore	Ignore	0	0
ps_reg_02	feature	region	interval	float64	Ignore	Ignore	0	0
ps_reg_03	feature	region	interval	float64	Ignore	Ignore	269456	269456(18
ps_car_01_cat	feature	car	nominal	int64	12	Medium Cardinality	267	267(0
ps_car_02_cat	feature	car	nominal	int64	2	Low Cardinality	10	10
ps_car_03_cat	feature	car	nominal	int64	2	Low Cardinality	1028142	1028142(69
ps_car_04_cat	feature	car	nominal	int64	10	Low Cardinality	0	0
ps_car_05_cat	feature	car	nominal	int64	2	Low Cardinality	666910	666910(44
ps_car_06_cat	feature	car	nominal	int64	18	Medium Cardinality	0	0

colname	role	category	level	dtype	cardinality	cardinality_level	missing	missing_p
ps_car_07_cat	feature	car	nominal	int64	2	Low Cardinality	28820	28820(1
ps_car_08_cat	feature	car	nominal	int64	2	Low Cardinality	0	0
ps_car_09_cat	feature	car	nominal	int64	5	Low Cardinality	1446	1446
ps_car_10_cat	feature	car	nominal	int64	3	Low Cardinality	0	0
ps_car_11_cat	feature	car	nominal	int64	104	High Cardinality	0	0
ps_car_11	feature	car	ordinal	int64	4	Low Cardinality	6	6
ps_car_12	feature	car	interval	float64	Ignore	Ignore	1	1
ps_car_13	feature	car	interval	float64	Ignore	Ignore	0	0
ps_car_14	feature	car	ratio	float64	Ignore	Ignore	106425	106425(7
ps_car_15	feature	car	interval	float64	Ignore	Ignore	0	0
ps_calc_01	feature	calculated	ratio	float64	Ignore	Ignore	0	0
ps_calc_02	feature	calculated	ratio	float64	Ignore	Ignore	0	0
ps_calc_03	feature	calculated	ratio	float64	Ignore	Ignore	0	0
ps_calc_04	feature	calculated	ordinal	int64	6	Low Cardinality	0	0
ps_calc_05	feature	calculated	ordinal	int64	7	Low Cardinality	0	0
ps_calc_06	feature	calculated	ordinal	int64	11	Medium Cardinality	0	0
ps_calc_07	feature	calculated	ordinal	int64	10	Low Cardinality	0	0
ps_calc_08	feature	calculated	ordinal	int64	12	Medium Cardinality	0	0
ps_calc_09	feature	calculated	ordinal	int64	8	Low Cardinality	0	0
ps_calc_10	feature	calculated	ordinal	int64	26	Medium Cardinality	0	0
ps_calc_11	feature	calculated	ordinal	int64	21	Medium Cardinality	0	0
ps_calc_12	feature	calculated	ordinal	int64	12	Medium Cardinality	0	0
ps_calc_13	feature	calculated	ordinal	int64	16	Medium Cardinality	0	0
ps_calc_14	feature	calculated	ordinal	int64	25	Medium Cardinality	0	0
ps_calc_15_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0
ps_calc_16_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0
ps_calc_17_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0
ps_calc_18_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0

colname	role	category	level	dtype	cardinality	cardinality_level	missing	missing_p
ps_calc_19_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0
ps_calc_20_bin	feature	calculated	binary	int64	2	Low Cardinality	0	0

In [6]:
```
metadata.groupby(['role','level']).size().reset_index(name = 'count')
```

Out[6]:

	role	level	count
0	feature	binary	17
1	feature	interval	5
2	feature	nominal	14
3	feature	ordinal	16
4	feature	ratio	5
5	id	nominal	1
6	target	binary	1

In [7]:
```
stats = fullset[metadata[metadata.dtype == 'float64'].index].describe()
stats
```

Out[7]:

	ps_reg_01	ps_reg_02	ps_reg_03	ps_car_12	ps_car_13	ps_car_14	
count	1.488028e+06	1.488028e+06	1.488028e+06	1.488028e+06	1.488028e+06	1.488028e+06	1.48
mean	6.110305e-01	4.395943e-01	5.514848e-01	3.799519e-01	8.134878e-01	2.763614e-01	3.0(
std	2.876763e-01	4.045123e-01	7.938159e-01	5.836187e-02	2.247024e-01	3.569623e-01	7.2
min	0.000000e+00	0.000000e+00	-1.000000e+00	-1.000000e+00	2.506191e-01	-1.000000e+00	0.0(
25%	4.000000e-01	2.000000e-01	5.250000e-01	3.162278e-01	6.710052e-01	3.339162e-01	2.8
50%	7.000000e-01	3.000000e-01	7.211103e-01	3.741657e-01	7.660406e-01	3.687818e-01	3.3
75%	9.000000e-01	6.000000e-01	1.001561e+00	4.000000e-01	9.061429e-01	3.964846e-01	3.6(
max	9.000000e-01	1.800000e+00	4.423517e+00	1.264911e+00	4.031301e+00	6.363961e-01	3.7

In [8]:
```
stats.columns[stats.loc['min'] == -1]
```

Out[8]:
```
Index(['ps_reg_03', 'ps_car_12', 'ps_car_14'], dtype='object')
```

In [9]:
```
stats.loc['std'].nsmallest(1).index[0]
```

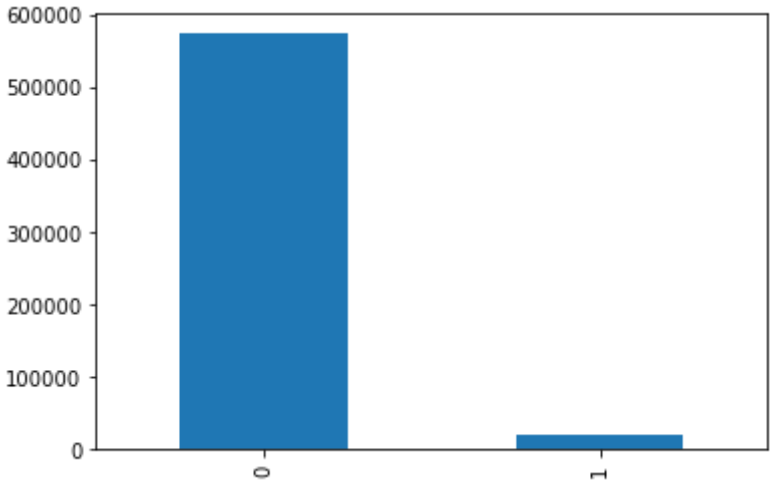
Out[9]:
```
'ps_car_12'
```

In [10]:
```
stats.columns[stats.loc['max'] > 4]
```

Out[10]:
```
Index(['ps_reg_03', 'ps_car_13'], dtype='object')
```

In [11]:
```
train.target.value_counts().plot(kind = 'bar')
```

Out[11]: <AxesSubplot:>



In [12]:
```
import pandas as pd
import numpy as np
import colorama
from colorama import Fore, Style
from tabulate import tabulate
```

In [13]:
```
def data_report(train,test,metadata,verbose = False):

    fullset = pd.concat([train,test]).reset_index(drop=True).fillna('未知')

    print(f"train总行数: {Fore.RED}{train.shape[0]}{Style.RESET_ALL} | test总
    print(f"train总列数: {Fore.RED}{train.shape[1]}{Style.RESET_ALL} | test总
    print(f"train总元素数: {train.size}")
    print(f"test总元素数: {test.size}")
    print('-'*50+ f"{Fore.RED}INFO{Style.RESET_ALL}"  + '-'*50)
    print('【train info】')
    train.info(verbose = verbose)
    print('-'*104)
    print('【test info】')
    test.info(verbose = verbose)

    if verbose:

        print('-'*48 + f"{Fore.RED}SUMMARY{Style.RESET_ALL}" + '-'*48)


        ############ SUMMARY #############
        print('*'*48 + f"{Fore.BLUE} COUNTS {Style.RESET_ALL}" + '*'*48)
        print('【Counts groupby role & level】'.upper())
        role_level_count = pd.DataFrame(
        {
            'count':metadata.groupby(['role','level']).size()
        }
        ).reset_index().sort_values(by = 'count',ascending=False)
        print(tabulate(role_level_count,tablefmt="grid",headers = ['role','

        print('【Counts groupby role & category】'.upper())
        role_cate_count = pd.DataFrame(
        {
            'count':metadata.groupby(['role','category']).size()
        }
        ).reset_index().sort_values(by = 'count',ascending=False)
        print(tabulate(role_cate_count,tablefmt="grid",headers = ['role','c

        print('【Counts groupby role & cardinality_level】'.upper())
        role_cardinality_count = pd.DataFrame(
        {
            'count':metadata.groupby(['role','cardinality_level']).size()
        }
        ).reset_index().sort_values(by = 'count',ascending=False)
        print(tabulate(role_cardinality_count,tablefmt="grid",headers = ['r


        print('*'*48 + f"{Fore.BLUE} MISSING {Style.RESET_ALL}" + '*'*48)
        print('【Cols to drop】'.upper())
        for col in metadata[metadata['keep'] == False].index:
            print(f" • {col}")

        print('【Cols to impute using (mode)】'.upper())
        for col in metadata[metadata['imputation'] == ('mode')].index:
            print(f" • {col}")
```

```
print(' 【Cols to impute using (mode|median)】 '.upper())
for col in metadata[metadata['imputation'] == ('mode','median')].in
    print(f" • {col}")

print(' 【Cols to impute using (mode|median|mean)】 '.upper())
for col in metadata[metadata['imputation'] == ('mode','median','mea
    print(f" • {col}")

print('*'*48 + f"{Fore.BLUE} CARDINALITY {Style.RESET_ALL}" + '*'*4
print(' 【Cols with medium cardinality】 ==> '.upper()+f'{Fore.YELLOW
for col in metadata[metadata['cardinality_level'] == 'Medium Cardin
    print(f" • {col}")

print(' 【Cols with High cardinality】 ==> '.upper()+f'{Fore.YELLOW}H
for col in metadata[metadata['cardinality_level'] == 'High Cardinal
    print(f" • {Fore.GREEN}{col}{Style.RESET_ALL}")


print('-'*42 + f"{Fore.RED}DESCRIPTIVE ANALYSIS{Style.RESET_ALL}" +
conti_descrip = fullset[metadata[metadata['level'].isin(['interval'
print(tabulate(conti_descrip.T,tablefmt="grid",headers = conti_desc

print('-'*50 + f"{Fore.RED}META{Style.RESET_ALL}" + '-'*50)
cols = ['role','category', 'level', 'dtype','cardinality', 'missing
print(tabulate(metadata[cols],tablefmt="grid",headers = cols))
```

In [14]:
```
data_report(train,test,metadata,verbose=True)
```

```
train总行数: 595212 | test总行数: 892816
train总列数: 59 | test总列数: 58
train总元素数: 35117508
test总元素数: 51783328
-----------------------------------------------------INFO-------------------
------------------------------
【train info】
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 595212 entries, 0 to 595211
Data columns (total 59 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   id              595212 non-null  int64  
 1   target          595212 non-null  int64  
 2   ps_ind_01       595212 non-null  int64  
 3   ps_ind_02_cat   595212 non-null  int64  
 4   ps_ind_03       595212 non-null  int64  
 5   ps_ind_04_cat   595212 non-null  int64  
 6   ps_ind_05_cat   595212 non-null  int64  
 7   ps_ind_06_bin   595212 non-null  int64  
 8   ps_ind_07_bin   595212 non-null  int64  
 9   ps_ind_08_bin   595212 non-null  int64  
 10  ps_ind_09_bin   595212 non-null  int64  
 11  ps_ind_10_bin   595212 non-null  int64  
 12  ps_ind_11_bin   595212 non-null  int64  
 13  ps_ind_12_bin   595212 non-null  int64  
 14  ps_ind_13_bin   595212 non-null  int64  
 15  ps_ind_14       595212 non-null  int64  
 16  ps_ind_15       595212 non-null  int64  
 17  ps_ind_16_bin   595212 non-null  int64  
 18  ps_ind_17_bin   595212 non-null  int64  
 19  ps_ind_18_bin   595212 non-null  int64  
 20  ps_reg_01       595212 non-null  float64
 21  ps_reg_02       595212 non-null  float64
 22  ps_reg_03       595212 non-null  float64
 23  ps_car_01_cat   595212 non-null  int64  
 24  ps_car_02_cat   595212 non-null  int64  
 25  ps_car_03_cat   595212 non-null  int64  
 26  ps_car_04_cat   595212 non-null  int64  
 27  ps_car_05_cat   595212 non-null  int64  
 28  ps_car_06_cat   595212 non-null  int64  
 29  ps_car_07_cat   595212 non-null  int64  
 30  ps_car_08_cat   595212 non-null  int64  
 31  ps_car_09_cat   595212 non-null  int64  
 32  ps_car_10_cat   595212 non-null  int64  
 33  ps_car_11_cat   595212 non-null  int64  
 34  ps_car_11       595212 non-null  int64  
 35  ps_car_12       595212 non-null  float64
 36  ps_car_13       595212 non-null  float64
 37  ps_car_14       595212 non-null  float64
 38  ps_car_15       595212 non-null  float64
 39  ps_calc_01      595212 non-null  float64
 40  ps_calc_02      595212 non-null  float64
 41  ps_calc_03      595212 non-null  float64
 42  ps_calc_04      595212 non-null  int64  
```

```
 43  ps_calc_05      595212 non-null  int64
 44  ps_calc_06      595212 non-null  int64
 45  ps_calc_07      595212 non-null  int64
 46  ps_calc_08      595212 non-null  int64
 47  ps_calc_09      595212 non-null  int64
 48  ps_calc_10      595212 non-null  int64
 49  ps_calc_11      595212 non-null  int64
 50  ps_calc_12      595212 non-null  int64
 51  ps_calc_13      595212 non-null  int64
 52  ps_calc_14      595212 non-null  int64
 53  ps_calc_15_bin  595212 non-null  int64
 54  ps_calc_16_bin  595212 non-null  int64
 55  ps_calc_17_bin  595212 non-null  int64
 56  ps_calc_18_bin  595212 non-null  int64
 57  ps_calc_19_bin  595212 non-null  int64
 58  ps_calc_20_bin  595212 non-null  int64
dtypes: float64(10), int64(49)
memory usage: 267.9 MB
--------------------------------------------------------------------------------
------------------------------
【test info】
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 892816 entries, 0 to 892815
Data columns (total 58 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   id              892816 non-null  int64
 1   ps_ind_01       892816 non-null  int64
 2   ps_ind_02_cat   892816 non-null  int64
 3   ps_ind_03       892816 non-null  int64
 4   ps_ind_04_cat   892816 non-null  int64
 5   ps_ind_05_cat   892816 non-null  int64
 6   ps_ind_06_bin   892816 non-null  int64
 7   ps_ind_07_bin   892816 non-null  int64
 8   ps_ind_08_bin   892816 non-null  int64
 9   ps_ind_09_bin   892816 non-null  int64
 10  ps_ind_10_bin   892816 non-null  int64
 11  ps_ind_11_bin   892816 non-null  int64
 12  ps_ind_12_bin   892816 non-null  int64
 13  ps_ind_13_bin   892816 non-null  int64
 14  ps_ind_14       892816 non-null  int64
 15  ps_ind_15       892816 non-null  int64
 16  ps_ind_16_bin   892816 non-null  int64
 17  ps_ind_17_bin   892816 non-null  int64
 18  ps_ind_18_bin   892816 non-null  int64
 19  ps_reg_01       892816 non-null  float64
 20  ps_reg_02       892816 non-null  float64
 21  ps_reg_03       892816 non-null  float64
 22  ps_car_01_cat   892816 non-null  int64
 23  ps_car_02_cat   892816 non-null  int64
 24  ps_car_03_cat   892816 non-null  int64
 25  ps_car_04_cat   892816 non-null  int64
 26  ps_car_05_cat   892816 non-null  int64
 27  ps_car_06_cat   892816 non-null  int64
 28  ps_car_07_cat   892816 non-null  int64
 29  ps_car_08_cat   892816 non-null  int64
 30  ps_car_09_cat   892816 non-null  int64
```

```
 31  ps_car_10_cat   892816 non-null  int64
 32  ps_car_11_cat   892816 non-null  int64
 33  ps_car_11       892816 non-null  int64
 34  ps_car_12       892816 non-null  float64
 35  ps_car_13       892816 non-null  float64
 36  ps_car_14       892816 non-null  float64
 37  ps_car_15       892816 non-null  float64
 38  ps_calc_01      892816 non-null  float64
 39  ps_calc_02      892816 non-null  float64
 40  ps_calc_03      892816 non-null  float64
 41  ps_calc_04      892816 non-null  int64
 42  ps_calc_05      892816 non-null  int64
 43  ps_calc_06      892816 non-null  int64
 44  ps_calc_07      892816 non-null  int64
 45  ps_calc_08      892816 non-null  int64
 46  ps_calc_09      892816 non-null  int64
 47  ps_calc_10      892816 non-null  int64
 48  ps_calc_11      892816 non-null  int64
 49  ps_calc_12      892816 non-null  int64
 50  ps_calc_13      892816 non-null  int64
 51  ps_calc_14      892816 non-null  int64
 52  ps_calc_15_bin  892816 non-null  int64
 53  ps_calc_16_bin  892816 non-null  int64
 54  ps_calc_17_bin  892816 non-null  int64
 55  ps_calc_18_bin  892816 non-null  int64
 56  ps_calc_19_bin  892816 non-null  int64
 57  ps_calc_20_bin  892816 non-null  int64
dtypes: float64(10), int64(48)
memory usage: 395.1 MB
-------------------------------------------------SUMMARY-----------------
------------------------------
********************************************** COUNTS *****************
*****************************
【COUNTS GROUPBY ROLE & LEVEL】
+----+---------+----------+---------+
|    | role    | level    |   count |
+====+=========+==========+=========+
|  0 | feature | binary   |      17 |
+----+---------+----------+---------+
|  3 | feature | ordinal  |      16 |
+----+---------+----------+---------+
|  2 | feature | nominal  |      14 |
+----+---------+----------+---------+
|  1 | feature | interval |       5 |
+----+---------+----------+---------+
|  4 | feature | ratio    |       5 |
+----+---------+----------+---------+
|  5 | id      | nominal  |       1 |
+----+---------+----------+---------+
|  6 | target  | binary   |       1 |
+----+---------+----------+---------+
【COUNTS GROUPBY ROLE & CATEGORY】
+----+---------+------------+---------+
|    | role    | category   |   count |
+====+=========+============+=========+
|  0 | feature | calculated |      20 |
+----+---------+------------+---------+
```

```
|  2 | feature | individual |      18 |
+----+---------+------------+---------+
|  1 | feature | car        |      16 |
+----+---------+------------+---------+
|  3 | feature | region     |       3 |
+----+---------+------------+---------+
|  4 | id      | other      |       1 |
+----+---------+------------+---------+
|  5 | target  | other      |       1 |
+----+---------+------------+---------+
【COUNTS GROUPBY ROLE & CARDINALITY_LEVEL】
+----+---------+--------------------+---------+
|    | role    | cardinality_level  |   count |
+====+=========+====================+=========+
|  2 | feature | Low Cardinality    |      35 |
+----+---------+--------------------+---------+
|  3 | feature | Medium Cardinality |      11 |
+----+---------+--------------------+---------+
|  1 | feature | Ignore             |      10 |
+----+---------+--------------------+---------+
|  0 | feature | High Cardinality   |       1 |
+----+---------+--------------------+---------+
|  4 | id      | Ignore             |       1 |
+----+---------+--------------------+---------+
|  5 | target  | Low Cardinality    |       1 |
+----+---------+--------------------+---------+
********************************************** MISSING ****************
******************************
【COLS TO DROP】
 • id
 • ps_car_03_cat
 • ps_car_05_cat
【COLS TO IMPUTE USING (MODE)】
 • ps_ind_02_cat
 • ps_ind_04_cat
 • ps_ind_05_cat
 • ps_car_01_cat
 • ps_car_02_cat
 • ps_car_07_cat
 • ps_car_09_cat
【COLS TO IMPUTE USING (MODE|MEDIAN)】
 • ps_car_11
【COLS TO IMPUTE USING (MODE|MEDIAN|MEAN)】
 • ps_reg_03
 • ps_car_12
 • ps_car_14
********************************************** CARDINALITY ************
**********************************
【COLS WITH MEDIUM CARDINALITY】 ==> PLEASE TAKE CARE OF USING ONEHOT-ENCO
DING
 • ps_ind_03
 • ps_ind_15
 • ps_car_01_cat
 • ps_car_06_cat
 • ps_calc_06
 • ps_calc_08
 • ps_calc_10
```

- ps_calc_11
- ps_calc_12
- ps_calc_13
- ps_calc_14

【COLS WITH HIGH CARDINALITY】 ==> PLEASE APPLY TARGET-ENCODING

- ps_car_11_cat

----------DESCRIPTIVE ANALYSIS----------

	count	mean	std	min	25%	50%	75%	max
ps_reg_01	1.48803e+06	0.611031	0.287676	0	0.4	0.7	0.9	0.9
ps_reg_02	1.48803e+06	0.439594	0.404512	0	0.2	0.3	0.6	1.8
ps_reg_03	1.48803e+06	0.551485	0.793816	-1	0.525	0.72111	1.00156	4.42352
ps_car_12	1.48803e+06	0.379952	0.0583619	-1	0.316228	0.374166	0.4	1.26491
ps_car_13	1.48803e+06	0.813488	0.224702	0.250619	0.671005	0.766041	0.906143	4.0313
ps_car_14	1.48803e+06	0.276361	0.356962	-1	0.333916	0.368782	0.396485	0.636396
ps_car_15	1.48803e+06	3.06735	0.729951	0	2.82843	3.31662	3.60555	3.74166
ps_calc_01	1.48803e+06	0.449682	0.287207	0	0.2	0.4	0.7	0.9
ps_calc_02	1.48803e+06	0.450107	0.287182	0	0.2	0.5	0.7	0.9
ps_calc_03	1.48803e+06	0.449972	0.287214	0	0.2	0.5	0.7	0.9

----------META----------

	role	category	level	dtype	cardinality	missing_percent	keep
id	id	other	nominal	int64	Ignore	0(0.0%)	False
target	target	other	binary	object	3	0(0.0%)	True
ps_ind_01	feature	individual	ordinal	int64	8	0(0.0%)	True
ps_ind_02_cat	feature	individual	nominal	int64	4	523(0.04%)	True
ps_ind_03	feature	individual	ordinal	int64	12	0(0.0%)	True
ps_ind_04_cat	feature	individual	nominal	int64	2	228(0.02%)	True
ps_ind_05_cat	feature	individual	nominal	int64	7	14519(0.98%)	True
ps_ind_06_bin	feature	individual	binary	int64	2	0(0.0%)	True
ps_ind_07_bin	feature	individual	binary	int64	2	0(0.0%)	True
ps_ind_08_bin	feature	individual	binary	int64	2	0(0.0%)	True
ps_ind_09_bin	feature	individual	binary	int64	2	0(0.0%)	True
ps_ind_10_bin	feature	individual	binary	int64	2	0(0.0%)	True
ps_ind_11_bin	feature	individual	binary	int64	2	0(0.0%)	True

```
----+------------------+--------+
| ps_ind_12_bin  | feature | individual | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_13_bin  | feature | individual | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_14      | feature | individual | ordinal  | int64   | 5
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_15      | feature | individual | ordinal  | int64   | 14
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_16_bin  | feature | individual | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_17_bin  | feature | individual | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_ind_18_bin  | feature | individual | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_reg_01      | feature | region     | ratio    | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_reg_02      | feature | region     | interval | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_reg_03      | feature | region     | interval | float64 | Ignore
| 269456(18.11%)    | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_01_cat  | feature | car        | nominal  | int64   | 12
| 267(0.02%)        | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_02_cat  | feature | car        | nominal  | int64   | 2
| 10(0.0%)          | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_03_cat  | feature | car        | nominal  | int64   | 2
| 1028142(69.09%)   | False  |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_04_cat  | feature | car        | nominal  | int64   | 10
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
```

```
| ps_car_05_cat  | feature | car        | nominal  | int64   | 2
| 666910(44.82%)    | False  |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_06_cat  | feature | car        | nominal  | int64   | 18
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_07_cat  | feature | car        | nominal  | int64   | 2
| 28820(1.94%)      | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_08_cat  | feature | car        | nominal  | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_09_cat  | feature | car        | nominal  | int64   | 5
| 1446(0.1%)        | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_10_cat  | feature | car        | nominal  | int64   | 3
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_11_cat  | feature | car        | nominal  | int64   | 104
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_11      | feature | car        | ordinal  | int64   | 4
| 6(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_12      | feature | car        | interval | float64 | Ignore
| 1(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_13      | feature | car        | interval | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_14      | feature | car        | ratio    | float64 | Ignore
| 106425(7.15%)     | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_car_15      | feature | car        | interval | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_01     | feature | calculated | ratio    | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_02     | feature | calculated | ratio    | float64 | Ignore
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_03     | feature | calculated | ratio    | float64 | Ignore
```

```
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_04     | feature | calculated | ordinal  | int64   | 6
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_05     | feature | calculated | ordinal  | int64   | 7
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_06     | feature | calculated | ordinal  | int64   | 11
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_07     | feature | calculated | ordinal  | int64   | 10
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_08     | feature | calculated | ordinal  | int64   | 12
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_09     | feature | calculated | ordinal  | int64   | 8
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_10     | feature | calculated | ordinal  | int64   | 26
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_11     | feature | calculated | ordinal  | int64   | 21
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_12     | feature | calculated | ordinal  | int64   | 12
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_13     | feature | calculated | ordinal  | int64   | 16
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_14     | feature | calculated | ordinal  | int64   | 25
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_15_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_16_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+------------------+--------+
| ps_calc_17_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
```

```
+----------------+---------+------------+----------+---------+-----------
----+-------------------+--------+
| ps_calc_18_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+-------------------+--------+
| ps_calc_19_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+-------------------+--------+
| ps_calc_20_bin | feature | calculated | binary   | int64   | 2
| 0(0.0%)           | True   |
+----------------+---------+------------+----------+---------+-----------
----+-------------------+--------+
```