In [1]:

```
import pandas as pd
import numpy as np
import os
```

In [3]:

```
print(os.getcwd())
os.chdir('D:/OneDrive/ASU/Humana_Case_Competition')
print(os.getcwd())
```

```
D:\OneDrive\ASU\Humana_Case_Competition
D:\OneDrive\ASU\Humana_Case_Competition
```

In [5]:

```
humana= pd.read_csv('train.csv')
```

```
D:\1DataAnalytics\Python\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3
058: DtypeWarning: Columns (80,193) have mixed types. Specify dtype option on import or
set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [15]:

```
holdout=pd.read_csv('test.csv')
```

```
D:\1DataAnalytics\Python\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3
058: DtypeWarning: Columns (79) have mixed types. Specify dtype option on import or set
low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [37]:

```
print("Training:",humana.shape,", Testing:", holdout.shape)
```

```
Training: (69572, 826) , Testing: (17681, 825)
```

In [23]:

```
percent_missing_humana = humana.isnull().sum() * 100 / len(humana)
missing_value_humana = pd.DataFrame({'percent_missing': percent_missing})
```

In [24]:
```
missing_value_humana.sort_values('percent_missing', inplace=True, ascending=False)
missing_value_humana
```

Out[24]:

	percent_missing
hedis_ami	99.665095
hedis_cmc_ldc_c_control	78.957052
hedis_cmc_ldc_c_screen	78.954177
cons_homstat	27.712298
cons_ret_y	27.710861
...	...
submcc_ano_dig_pmpm_ct	0.000000
submcc_ano_gu_pmpm_ct	0.000000
submcc_ano_hrt_pmpm_ct	0.000000
submcc_ano_mus_pmpm_ct	0.000000
submcc_rsk_chol_ind	0.000000

826 rows × 1 columns

In [25]:
```
percent_missing_holdout = holdout.isnull().sum() * 100 / len(holdout)
missing_value_holdout = pd.DataFrame({'percent_missing': percent_missing_holdout})
missing_value_holdout.sort_values('percent_missing', inplace=True, ascending=False)
missing_value_holdout
```

Out[25]:

	percent_missing
hedis_ami	99.666308
hedis_cmc_ldc_c_control	78.242181
hedis_cmc_ldc_c_screen	78.242181
cons_hcaccprf_h	27.085572
cons_retail_buyer	27.085572
...	...
submcc_ano_dig_pmpm_ct	0.000000
submcc_ano_gu_pmpm_ct	0.000000
submcc_ano_hrt_pmpm_ct	0.000000
submcc_ano_mus_pmpm_ct	0.000000
submcc_rsk_chol_ind	0.000000

825 rows × 1 columns

Combine two tables

In [47]:
```
#pd.get_dummies(holdout.drop(['person_id_syn'], axis=1))        1867
#pd.get_dummies(humana.drop(['person_id_syn'], axis=1))         1874
#holdout.insert(loc=1, column='transportation_issues', value=2)
```

In [49]:
```
frames=[humana, holdout]
fulldata = pd.concat(frames)
```

In [50]:
```
fulldata.shape
```

Out[50]:
```
(87253, 826)
```

Get Dummy Variables

In [51]:
```
Full_Dummy=pd.get_dummies(fulldata.drop(['person_id_syn'], axis=1))
```

In [52]:
```
Full_Dummy.index = fulldata.person_id_syn
```

In [53]:
```
Full_Dummy.head()
```

Out[53]:

	transportation_issues	est_age	smoker_current_ind	smoker_former_ind
person_id_syn				
0002MOb79ST17bLYAe46eIc2	0	62	1	0
0004cMOS6bTLf34Y7AIca8f3	0	59	1	0
000536M9O3ST98LaYaeA29Ia	1	63	0	0
0009bMO9SfTLYe77A51l4ac3	0	75	0	0
000M7OeS66bTL8bY89Aa16Ie	0	51	1	0

5 rows × 1874 columns

In [55]:
```
Train = Full_Dummy.iloc[0:69572, :]
Test = Full_Dummy.iloc[69572:, :]
```

In [58]:
```
print("Training:",Train.shape,", Testing:", Test.shape)
```

```
Training: (69572, 1874) , Testing: (17681, 1874)
```

In [60]:
```
#Train.to_csv('Train_Dummy.csv')
#Test.to_csv('Test_Dummy.csv')
```

In [ ]:
```

```