In [1]:
```
# Supress Warnings
import warnings
warnings.filterwarnings('ignore')

# Importing librarieshttp://localhost:8888/notebooks/project/preprocessing.ipynb#
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# visulaisation
from matplotlib.pyplot import xticks
%matplotlib inline

# Data display coustomization
#pd.set_option('display.max_rows', 50)
#pd.set_option('display.max_columns', 50)
```

In [2]:
```
df = pd.read_csv('2021_Competition_Training.csv')
df.head(5)
```

Out[2]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacp
0	0	1MObcfaSTac85Lca0Y8bbA6I	0	0.000000	7.910346	0.049
1	1	5M89OSTL580dYeA849d3480I	0	0.000000	1.730272	0.095
2	2	MdOS23TLe18Y60043Acfa2I9	0	0.000000	5.015501	0.022
3	3	2ccMO510abSaT79cLfaYAIe4	0	2.266667	4.049586	0.070
4	4	0M9811Ocb1ST94LY3f5A9I00	0	0.000000	0.618606	0.074

5 rows × 368 columns

In [3]:
```
df.shape
```

Out[3]:
```
(974842, 368)
```

In [4]:
```
#giving trining and test data
df_copy = df.copy()
train = df_copy.drop(columns=['covid_vaccination'])
test = df[['covid_vaccination']]
train.head()
```

Out[4]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacp
0	0	1MObcfaSTac85Lca0Y8bbA6I	0	0.000000	7.910346	0.049
1	1	5M89OSTL580dYeA849d3480I	0	0.000000	1.730272	0.095
2	2	MdOS23TLe18Y60043Acfa2I9	0	0.000000	5.015501	0.022
3	3	2ccMO510abSaT79cLfaYAIe4	0	2.266667	4.049586	0.070
4	4	0M9811Ocb1ST94LY3f5A9I00	0	0.000000	0.618606	0.074

5 rows × 367 columns

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Unnamed: 0', 'ID', 'auth_3mth_post_acute_dia',
       'rx_gpi2_72_pmpm_cost_6to9m_b4', 'atlas_pct_laccess_child15',
       'atlas_recfacpth14', 'atlas_pct_fmrkt_frveg16',
       'atlas_pct_free_lunch14', 'bh_ip_snf_net_paid_pmpm_cost_9to12m_b4',
       'auth_3mth_acute_ckd',
       ...
       'auth_3mth_post_acute_end', 'auth_3mth_acute_mus',
       'atlas_perpov_1980_0711', 'atlas_pct_laccess_white15',
       'auth_3mth_post_acute_mean_los', 'rx_gpi2_66_pmpm_ct',
       'auth_3mth_acute_gus', 'rx_generic_dist_gpi6_pmpm_ct_t_9-6-3m_b4',
       'atlas_low_education_2015_update', 'race_cd'],
      dtype='object', length=368)
```

In [6]:
```
#Saving missing values in a variable
a = train.isnull().sum()/len(df)*100
variables = train.columns
# saving column names in a variable
variables_no_missing=[]
variables_missing=[]
for i in range(len(variables)):
    if a[i]<=70:
        variables_no_missing.append(variables[i])
    else:
        variables_missing.append(variables[i])
```

In [7]:
```
variables_no_missing
```

Out[7]:
```
['Unnamed: 0',
 'ID',
 'auth_3mth_post_acute_dia',
 'rx_gpi2_72_pmpm_cost_6to9m_b4',
 'atlas_pct_laccess_child15',
 'atlas_recfacpth14',
 'atlas_pct_fmrkt_frveg16',
 'atlas_pct_free_lunch14',
 'bh_ip_snf_net_paid_pmpm_cost_9to12m_b4',
 'auth_3mth_acute_ckd',
 'bh_ncal_pmpm_ct',
 'src_div_id',
 'total_bh_copay_pmpm_cost_t_9-6-3m_b4',
 'bh_ip_snf_net_paid_pmpm_cost_3to6m_b4',
 'cons_chmi',
 'mcc_ano_pmpm_ct_t_9-6-3m_b4',
 'auth_3mth_post_acute_trm',
 'rx_maint_pmpm_cost_t_12-9-6m_b4',
 'auth_3mth_post_acute_rsk',
 'cons_ltmedicr',
 'rx_gpi4_6110_pmpm_ct',
 'atlas_pc_snapben15',
 'credit_bal_nonmtgcredit_60dpd',
 'rx_bh_mbr_resp_pmpm_cost_9to12m_b4',
 'rx_nonbh_pmpm_cost_t_9-6-3m_b4',
 'atlas_pct_laccess_nhna15',
 'auth_3mth_acute_vco',
 'credit_hh_nonmtgcredit_60dpd',
 'rx_bh_pmpm_ct_0to3m_b4',
 'auth_3mth_dc_ltac',
 'cons_lwcm10',
 'auth_3mth_post_acute_inj',
 'atlas_fsrpth14',
 'auth_3mth_dc_home',
 'atlas_wicspth12',
 'rx_gpi2_17_pmpm_cost_t_12-9-6m_b4',
 'cons_hxmioc',
 'rx_generic_pmpm_cost_t_6-3-0m_b4',
 'cmsd2_sns_digest_abdomen_pmpm_ct',
 'atlas_ghveg_farms12',
 'credit_hh_bankcardcredit_60dpd',
 'total_outpatient_allowed_pmpm_cost_6to9m_b4',
 'cons_cwht',
 'atlas_netmigrationrate1016',
 'atlas_pct_laccess_snap15',
 'bh_ncdm_ind',
 'rx_nonmaint_mbr_resp_pmpm_cost_9to12m_b4',
 'atlas_retirement_destination_2015_upda',
 'rx_overall_mbr_resp_pmpm_cost_t_6-3-0m_b4',
 'atlas_naturalchangerate1016',
 'ccsp_236_pct',
 'bh_ip_snf_mbr_resp_pmpm_cost_6to9m_b4',
 'rx_overall_dist_gpi6_pmpm_ct_t_6-3-0m_b4',
 'auth_3mth_post_acute_ben',
 'atlas_pct_laccess_hisp15',
 'auth_3mth_dc_no_ref',
 'rx_overall_mbr_resp_pmpm_cost',
 'rx_overall_gpi_pmpm_ct_0to3m_b4',
 'auth_3mth_dc_snf',
 'rx_phar_cat_humana_pmpm_ct_t_9-6-3m_b4',
 'atlas_pct_laccess_hhnv15',
 'auth_3mth_acute_ccs_048',
 'bh_ip_snf_net_paid_pmpm_cost_0to3m_b4',
 'auth_3mth_acute_end',
 'auth_3mth_psychic',
 'atlas_hiamenity',
 'auth_3mth_bh_acute',
 'credit_bal_consumerfinance',
 'auth_3mth_acute_chf',
 'rx_overall_gpi_pmpm_ct_t_6-3-0m_b4',
 'rwjf_uninsured_pct',
 'mcc_chf_pmpm_ct_t_9-6-3m_b4',
```

```
'rx_mail_mbr_resp_pmpm_cost_0to3m_b4',
'bh_urgent_care_copay_pmpm_cost_t_12-9-6m_b4',
'auth_3mth_hospice',
'auth_3mth_acute_bld',
'atlas_pct_wic15',
'ccsp_193_pct',
'auth_3mth_dc_hospice',
'auth_3mth_acute_ccs_030',
'atlas_pct_fmrkt_baked16',
'rx_nonmaint_mbr_resp_pmpm_cost',
'auth_3mth_acute_skn',
'atlas_veg_farms12',
'atlas_vlfoodsec_13_15',
'rx_gpi2_34_dist_gpi6_pmpm_ct',
'bh_ip_snf_net_paid_pmpm_cost',
'credit_hh_bankcard_severederog',
'rx_hum_16_pmpm_ct',
'est_age',
'rx_maint_pmpm_cost_t_6-3-0m_b4',
'cnt_cp_webstatement_pmpm_ct',
'atlas_pct_laccess_seniors15',
'phy_em_px_pct',
'atlas_percapitainc',
'rwjf_uninsured_adults_pct',
'rx_generic_mbr_resp_pmpm_cost_0to3m_b4',
'auth_3mth_acute_neo',
'rwjf_air_pollute_density',
'rx_gpi2_02_pmpm_cost',
'atlas_recfac14',
'cons_mobplus',
'lab_albumin_loinc_pmpm_ct',
'atlas_pct_obese_adults13',
'rx_maint_net_paid_pmpm_cost_t_12-9-6m_b4',
'rev_pm_obsrm_pmpm_ct',
'atlas_pct_sfsp15',
'total_physician_office_net_paid_pmpm_cost_9to12m_b4',
'atlas_pc_dirsales12',
'med_ip_snf_admit_days_pmpm',
'rej_med_outpatient_visit_ct_pmpm_t_6-3-0m_b4',
'auth_3mth_post_acute_vco',
'cms_tot_partd_payment_amt',
'rx_nonotc_dist_gpi6_pmpm_ct',
'rx_nonmaint_pmpm_ct',
'rx_nonbh_mbr_resp_pmpm_cost_6to9m_b4',
'cons_stlnindx',
'atlas_hipov_1115',
'auth_3mth_post_acute_dig',
'rx_nonbh_mbr_resp_pmpm_cost',
'atlas_redemp_snaps16',
'atlas_berry_farms12',
'rej_med_ip_snf_coins_pmpm_cost_t_9-6-3m_b4',
'rwjf_inactivity_pct',
'rx_gpi2_72_pmpm_ct_6to9m_b4',
'cons_n2pmr',
'med_physician_office_allowed_pmpm_cost_t_9-6-3m_b4',
'auth_3mth_acute_res',
'rev_cms_ct_pmpm_ct',
'atlas_foodhub16',
'total_physician_office_copay_pmpm_cost',
'auth_3mth_acute_dig',
'auth_3mth_dc_acute_rehab',
'atlas_pct_fmrkt_anmlprod16',
'auth_3mth_post_acute_hdz',
'bh_ip_snf_mbr_resp_pmpm_cost_3to6m_b4',
'auth_3mth_acute_ccs_172',
'credit_num_agencyfirstmtg',
'total_physician_office_net_paid_pmpm_cost_t_9-6-3m_b4',
'auth_3mth_acute_ccs_154',
'atlas_type_2015_mining_no',
'atlas_agritrsm_rct12',
'rx_days_since_last_script',
'atlas_pct_laccess_pop15',
'auth_3mth_post_acute_res',
'auth_3mth_acute_inf',
'rx_gpi2_01_pmpm_cost_0to3m_b4',
'atlas_povertyallagespct',
'rwjf_uninsured_child_pct',
'rx_branded_pmpm_ct_t_6-3-0m_b4',
'med_outpatient_deduct_pmpm_cost_t_9-6-3m_b4',
'credit_bal_mtgcredit_new',
'atlas_low_employment_2015_update',
'atlas_pct_diabetes_adults13',
'atlas_pct_laccess_nhasian15',
'atlas_deep_pov_all',
'atlas_net_international_migration_rate',
'atlas_deep_pov_children',
'bh_ncdm_pct',
'auth_3mth_non_er',
'atlas_foodinsec_child_03_11',
```

```
'rx_branded_mbr_resp_pmpm_cost',
'atlas_pc_wic_redemp12',
'rwjf_mv_deaths_rate',
'auth_3mth_acute_cad',
'atlas_pct_reduced_lunch14',
'cons_nwperadult',
'total_allowed_pmpm_cost_t_9-6-3m_b4',
'rx_hum_28_pmpm_cost',
'mabh_seg',
'cms_orig_reas_entitle_cd',
'atlas_totalocchu',
'med_physician_office_ds_clm_6to9m_b4',
'bh_ip_snf_mbr_resp_pmpm_cost_9to12m_b4',
'atlas_pct_loclfarm12',
'rx_generic_mbr_resp_pmpm_cost',
'total_outpatient_mbr_resp_pmpm_cost_6to9m_b4',
'auth_3mth_post_acute_cir',
'rx_gpi4_3400_pmpm_ct',
'auth_3mth_post_acute_cer',
'lab_dist_loinc_pmpm_ct',
'atlas_pct_nslp15',
'rx_generic_pmpm_ct_0to3m_b4',
'oontwk_mbr_resp_pmpm_cost_t_6-3-0m_b4',
'atlas_pct_laccess_lowi15',
'bh_ncal_ind',
'auth_3mth_post_acute_mus',
'atlas_pct_fmrkt_sfmnp16',
'hum_region',
'rx_nonmail_dist_gpi6_pmpm_ct_t_9-6-3m_b4',
'atlas_pct_loclsale12',
'bh_ip_snf_net_paid_pmpm_cost_6to9m_b4',
'rej_med_er_net_paid_pmpm_cost_t_9-6-3m_b4',
'credit_bal_autobank',
'med_outpatient_mbr_resp_pmpm_cost_t_9-6-3m_b4',
'rx_overall_mbr_resp_pmpm_cost_0to3m_b4',
'rx_tier_2_pmpm_ct_3to6m_b4',
'rx_nonbh_net_paid_pmpm_cost',
'rx_maint_pmpm_ct_9to12m_b4',
'rx_nonbh_net_paid_pmpm_cost_t_6-3-0m_b4',
'atlas_type_2015_recreation_no',
'auth_3mth_post_acute_sns',
'rx_gpi2_39_pmpm_cost_t_6-3-0m_b4',
'atlas_type_2015_update',
'cms_risk_adjustment_factor_a_amt',
'total_ip_maternity_net_paid_pmpm_cost_t_12-9-6m_b4',
'rx_generic_pmpm_cost',
'cmsd2_eye_retina_pmpm_ct',
'auth_3mth_acute_can',
'auth_3mth_post_acute',
'auth_3mth_facility',
'rx_days_since_last_script_0to3m_b4',
'atlas_population_loss_2015_update',
'rx_maint_pmpm_ct_t_6-3-0m_b4',
'auth_3mth_post_acute_men',
'auth_3mth_acute_mean_los',
'credit_num_autofinance',
'cons_rxmaint',
'rx_mail_net_paid_pmpm_cost_t_6-3-0m_b4',
'auth_3mth_home',
'rx_maint_mbr_resp_pmpm_cost_6to9m_b4',
'cons_hxwearbl',
'total_physician_office_mbr_resp_pmpm_cost_t_9-6-3m_b4',
'atlas_pct_laccess_black15',
'atlas_hh65plusalonepct',
'atlas_farm_to_school13',
'auth_3mth_acute_inj',
'auth_3mth_acute_ccs_153',
'rej_days_since_last_clm',
'auth_3mth_transplant',
'bh_outpatient_net_paid_pmpm_cost',
'atlas_dirsales_farms12',
'rx_generic_pmpm_cost_6to9m_b4',
'rev_cms_ansth_pmpm_ct',
'atlas_convspth14',
'total_med_allowed_pmpm_cost_9to12m_b4',
'rx_mail_mbr_resp_pmpm_cost_t_9-6-3m_b4',
'med_outpatient_visit_ct_pmpm_t_12-9-6m_b4',
'rx_nonbh_pmpm_ct_t_9-6-3m_b4',
'auth_3mth_acute',
'rx_nonbh_pmpm_ct_0to3m_b4',
'atlas_pc_ffrsales12',
'auth_3mth_dc_left_ama',
'credit_bal_bankcard_severederog',
'atlas_povertyunder18pct',
'rx_tier_1_pmpm_ct_0to3m_b4',
'auth_3mth_acute_ccs_227',
'cons_estinv30_rc',
'auth_3mth_bh_acute_men',
'rx_gpi2_34_pmpm_ct',
```

```
'auth_3mth_dc_custodial',
'atlas_veg_acrespth12',
'atlas_grocpth14',
'total_med_net_paid_pmpm_cost_t_6-3-0m_b4',
'rx_gpi2_90_dist_gpi6_pmpm_ct_9to12m_b4',
'atlas_csa12',
'sex_cd',
'rx_gpi2_62_pmpm_cost_t_9-6-3m_b4',
'rx_overall_gpi_pmpm_ct_t_12-9-6m_b4',
'auth_3mth_ltac',
'cons_hhcomp',
'auth_3mth_acute_hdz',
'cons_rxadhs',
'auth_3mth_acute_men',
'atlas_pct_fmrkt_snap16',
'met_obe_diag_pct',
'cms_partd_ra_factor_amt',
'atlas_pct_sbp15',
'rwjf_resident_seg_black_inx',
'atlas_pct_cacfp15',
'auth_3mth_rehab',
'pdc_lip',
'atlas_ffrpth14',
'credit_num_autobank_new',
'auth_3mth_acute_ccs_086',
'rx_tier_2_pmpm_ct',
'cons_n2pwh',
'rx_nonmaint_dist_gpi6_pmpm_ct_t_12-9-6m_b4',
'atlas_berry_acrespth12',
'atlas_pct_fmrkt_credit16',
'atlas_slhouse12',
'atlas_pc_fsrsales12',
'credit_hh_1stmtgcredit',
'auth_3mth_snf_post_hsp',
'atlas_pct_fmrkt_wiccash16',
'atlas_foodinsec_13_15',
'auth_3mth_acute_cer',
'cons_rxadhm',
'atlas_fmrktpth16',
'rx_nonotc_pmpm_cost_t_6-3-0m_b4',
'cci_dia_m_pmpm_ct',
'auth_3mth_acute_trm',
'cons_n2phi',
'bh_physician_office_copay_pmpm_cost_6to9m_b4',
'rwjf_income_inequ_ratio',
'rej_total_physician_office_visit_ct_pmpm_0to3m_b4',
'auth_3mth_acute_dia',
'credit_num_nonmtgcredit_60dpd',
'auth_3mth_snf_direct',
'credit_bal_autofinance_new',
'auth_3mth_acute_ccs_067',
'auth_3mth_acute_ccs_043',
'rwjf_men_hlth_prov_ratio',
'auth_3mth_dc_home_health',
'rx_gpi2_56_dist_gpi6_pmpm_ct_3to6m_b4',
'cmsd2_sns_genitourinary_pmpm_ct',
'auth_3mth_acute_cir',
'auth_3mth_acute_ner',
'auth_3mth_acute_ccs_094',
'med_ambulance_coins_pmpm_cost_t_9-6-3m_b4',
'hedis_dia_hba1c_ge9',
'bh_ncal_pct',
'atlas_pct_snap16',
'ccsp_227_pct',
'atlas_ghveg_sqftpth12',
'rx_days_since_last_script_6to9m_b4',
'atlas_orchard_acrespth12',
'atlas_persistentchildpoverty_1980_2011',
'auth_3mth_post_acute_cad',
'atlas_pct_laccess_multir15',
'cons_cgqs',
'ccsp_065_pmpm_ct',
'auth_3mth_acute_ccs_044',
'atlas_medhhinc',
'rx_maint_net_paid_pmpm_cost_t_9-6-3m_b4',
'rwjf_mental_distress_pct',
'bh_ip_snf_admit_days_pmpm_t_9-6-3m_b4',
'rx_phar_cat_cvs_pmpm_ct_t_9-6-3m_b4',
'zip_cd',
'auth_3mth_post_acute_ckd',
'atlas_pct_laccess_nhpi15',
'auth_3mth_post_acute_ner',
'auth_3mth_post_er',
'credit_num_consumerfinance_new',
'rx_gpi2_49_pmpm_cost_0to3m_b4',
'cons_chva',
'atlas_avghhsize',
'rx_overall_net_paid_pmpm_cost_6to9m_b4',
'atlas_ownhomepct',
```

```
'atlas_orchard_farms12',
'total_physician_office_visit_ct_pmpm_t_6-3-0m_b4',
'atlas_pct_fmrkt_wic16',
'rx_gpi2_33_pmpm_ct_0to3m_b4',
'auth_3mth_post_acute_chf',
'rwjf_social_associate_rate',
'atlas_freshveg_farms12',
'auth_3mth_acute_ccs_042',
'auth_3mth_post_acute_inf',
'auth_3mth_acute_sns',
'days_since_last_clm_0to3m_b4',
'auth_3mth_dc_other',
'auth_3mth_bh_acute_mean_los',
'mcc_end_pct',
'auth_3mth_post_acute_gus',
'cons_lwcm07',
'atlas_pct_fmrkt_otherfood16',
'auth_3mth_post_acute_end',
'auth_3mth_acute_mus',
'atlas_perpov_1980_0711',
'atlas_pct_laccess_white15',
'auth_3mth_post_acute_mean_los',
'rx_gpi2_66_pmpm_ct',
'auth_3mth_acute_gus',
'rx_generic_dist_gpi6_pmpm_ct_t_9-6-3m_b4',
'atlas_low_education_2015_update',
'race_cd']
```

In [8]:
```
variables_missing
```

Out[8]:
```
['lang_spoken_cd']
```

In [9]:
```
len(variables_no_missing)
```

Out[9]:
```
366
```

In [10]:
```
train.drop(['lang_spoken_cd'], axis=1, inplace=True)
train.head()
```

Out[10]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacp
0	0	1MObcfaSTac85Lca0Y8bbA6I	0	0.000000	7.910346	0.049
1	1	5M89OSTL580dYeA849d3480I	0	0.000000	1.730272	0.095
2	2	MdOS23TLe18Y60043Acfa2I9	0	0.000000	5.015501	0.022
3	3	2ccMO510abSaT79cLfaYAIe4	0	2.266667	4.049586	0.070
4	4	0M9811Ocb1ST94LY3f5A9I00	0	0.000000	0.618606	0.074

5 rows × 366 columns

In [11]:
```
#Imputation with mode value for the categorical variables
for col in variables_no_missing:
    train[col].fillna(train[col].mode()[0], inplace=True)
```

In [12]:
```
#categorical Columns that contain object data type
categ_cols = train.select_dtypes(include='object')
categ_cols
```

Out[12]:

	ID	auth_3mth_post_acute_dia	bh_ip_snf_net_paid_pmpm_cost_9to12m_b4	auth_3mth_acute_ckd	src_div_id	to
0	1MObcfaSTac85Lca0Y8bbA6I	0	0.0	0	000	
1	5M89OSTL580dYeA849d3480I	0	0.0	0	000	
2	MdOS23TLe18Y60043Acfa2I9	0	0.0	0	000	
3	2ccMO510abSaT79cLfaYAIe4	0	0.0	0	000	
4	0M9811Ocb1ST94LY3f5A9I00	0	0.0	0	000	
...	...	...	...	...	...	
974837	M047fa5OSffe1T8L9cYAI5f9	0	0.0	0	001	

974838	Md52O4STLY3A3e3bdI1f449f	0	0.0	0	000
974839	7M86Oe06dde42STLa0Y7AbI4	0	0.0	0	000
974840	f96bMfO720ca93S4T5LY4AI8	0	0.0	0	000
974841	MOS171T1cLa79afYe24Ad1Ic	0	0.0	0	001

974842 rows × 113 columns

In [13]:
```
#Printing cardinality of each categorical column
categ_cols.nunique()
```

Out[13]:
```
ID                                          974842
auth_3mth_post_acute_dia                         3
bh_ip_snf_net_paid_pmpm_cost_9to12m_b4           3
auth_3mth_acute_ckd                              3
src_div_id                                      14
                                              ...
auth_3mth_dc_other                               4
auth_3mth_bh_acute_mean_los                      5
auth_3mth_post_acute_gus                         3
auth_3mth_acute_mus                              4
rx_generic_dist_gpi6_pmpm_ct_t_9-6-3m_b4        12
Length: 113, dtype: int64
```

In [14]:
```
categ_cols.columns
```

Out[14]:
```
Index(['ID', 'auth_3mth_post_acute_dia',
       'bh_ip_snf_net_paid_pmpm_cost_9to12m_b4', 'auth_3mth_acute_ckd',
       'src_div_id', 'total_bh_copay_pmpm_cost_t_9-6-3m_b4',
       'bh_ip_snf_net_paid_pmpm_cost_3to6m_b4', 'mcc_ano_pmpm_ct_t_9-6-3m_b4',
       'auth_3mth_post_acute_trm', 'rx_maint_pmpm_cost_t_12-9-6m_b4',
       ...
       'rx_phar_cat_cvs_pmpm_ct_t_9-6-3m_b4', 'auth_3mth_post_er',
       'total_physician_office_visit_ct_pmpm_t_6-3-0m_b4',
       'rx_gpi2_33_pmpm_ct_0to3m_b4', 'auth_3mth_post_acute_chf',
       'auth_3mth_dc_other', 'auth_3mth_bh_acute_mean_los',
       'auth_3mth_post_acute_gus', 'auth_3mth_acute_mus',
       'rx_generic_dist_gpi6_pmpm_ct_t_9-6-3m_b4'],
      dtype='object', length=113)
```

In [15]:
```
train['auth_3mth_post_acute_dia'] = train['auth_3mth_post_acute_dia'].astype(str)
```

In [16]:
```
train['auth_3mth_post_acute_dia']
```

Out[16]:
```
0         0
1         0
2         0
3         0
4         0
         ..
974837    0
974838    0
974839    0
974840    0
974841    0
Name: auth_3mth_post_acute_dia, Length: 974842, dtype: object
```

In [17]:
```
from sklearn.preprocessing import LabelEncoder

train['auth_3mth_post_acute_dia'] = LabelEncoder().fit_transform(train['auth_3mth_post_acute_dia'])
```

In [18]:
```
train['auth_3mth_post_acute_dia'].head(10220)
```

Out[18]:
```
0        1
1        1
2        1
3        1
4        1
        ..
10215    1
10216    1
```

```
10217    1
10218    1
10219    1
Name: auth_3mth_post_acute_dia, Length: 10220, dtype: int32
```

In [19]:

```
list1 = ['ID', 'auth_3mth_post_acute_dia',
       'bh_ip_snf_net_paid_pmpm_cost_9to12m_b4', 'auth_3mth_acute_ckd',
       'src_div_id', 'total_bh_copay_pmpm_cost_t_9-6-3m_b4',
       'bh_ip_snf_net_paid_pmpm_cost_3to6m_b4', 'mcc_ano_pmpm_ct_t_9-6-3m_b4',
       'auth_3mth_post_acute_trm', 'rx_maint_pmpm_cost_t_12-9-6m_b4']

for col in list1:
    train[col] = train[col].astype(str)

for col in list1:
    train[col] = LabelEncoder().fit_transform(train[col])
```

In [23]:

```
list2 = ['rx_phar_cat_cvs_pmpm_ct_t_9-6-3m_b4', 'auth_3mth_post_er',
       'total_physician_office_visit_ct_pmpm_t_6-3-0m_b4',
       'rx_gpi2_33_pmpm_ct_0to3m_b4', 'auth_3mth_post_acute_chf',
       'auth_3mth_dc_other', 'auth_3mth_bh_acute_mean_los',
       'auth_3mth_post_acute_gus', 'auth_3mth_acute_mus',
       'rx_generic_dist_gpi6_pmpm_ct_t_9-6-3m_b4']

for col in list2:
    train[col] = train[col].astype(str)

for col in list2:
    train[col] = LabelEncoder().fit_transform(train[col])
```

In [24]:

```
#categorical Columns
categ_cols = train.select_dtypes(include='object')
categ_cols.columns
```

Out[24]:

```
Index(['bh_ip_snf_mbr_resp_pmpm_cost_9to12m_b4', 'auth_3mth_post_acute_cer',
       'oontwk_mbr_resp_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_post_acute_mus',
       'hum_region', 'rx_nonmail_dist_gpi6_pmpm_ct_t_9-6-3m_b4',
       'bh_ip_snf_net_paid_pmpm_cost_6to9m_b4',
       'rej_med_er_net_paid_pmpm_cost_t_9-6-3m_b4',
       'med_outpatient_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'rx_nonbh_net_paid_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_post_acute_sns',
       'rx_gpi2_39_pmpm_cost_t_6-3-0m_b4',
       'total_ip_maternity_net_paid_pmpm_cost_t_12-9-6m_b4',
       'auth_3mth_acute_can', 'auth_3mth_post_acute', 'auth_3mth_facility',
       'rx_maint_pmpm_ct_t_6-3-0m_b4', 'auth_3mth_post_acute_men',
       'rx_mail_net_paid_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_home',
       'total_physician_office_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'auth_3mth_transplant', 'rev_cms_ansth_pmpm_ct',
       'rx_mail_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'med_outpatient_visit_ct_pmpm_t_12-9-6m_b4',
       'rx_nonbh_pmpm_ct_t_9-6-3m_b4', 'auth_3mth_acute',
       'auth_3mth_dc_left_ama', 'auth_3mth_acute_ccs_227',
       'auth_3mth_dc_custodial', 'total_med_net_paid_pmpm_cost_t_6-3-0m_b4',
       'rx_gpi2_90_dist_gpi6_pmpm_ct_9to12m_b4', 'sex_cd',
       'rx_gpi2_62_pmpm_cost_t_9-6-3m_b4',
       'rx_overall_gpi_pmpm_ct_t_12-9-6m_b4', 'auth_3mth_ltac', 'cons_hhcomp',
       'rx_nonmaint_dist_gpi6_pmpm_ct_t_12-9-6m_b4', 'auth_3mth_snf_post_hsp',
       'rx_nonotc_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_acute_trm',
       'rej_total_physician_office_visit_ct_pmpm_0to3m_b4',
       'auth_3mth_snf_direct', 'auth_3mth_dc_home_health',
       'rx_gpi2_56_dist_gpi6_pmpm_ct_3to6m_b4', 'auth_3mth_acute_ner',
       'med_ambulance_coins_pmpm_cost_t_9-6-3m_b4', 'hedis_dia_hba1c_ge9',
       'ccsp_065_pmpm_ct', 'rx_maint_net_paid_pmpm_cost_t_9-6-3m_b4',
       'bh_ip_snf_admit_days_pmpm_t_9-6-3m_b4'],
      dtype='object')
```

In [25]:

```
list3 = ['rx_gpi4_6110_pmpm_ct', 'rx_nonbh_pmpm_cost_t_9-6-3m_b4',
       'auth_3mth_acute_vco', 'rx_bh_pmpm_ct_0to3m_b4', 'auth_3mth_dc_ltac',
       'auth_3mth_post_acute_inj', 'auth_3mth_dc_home',
       'rx_gpi2_17_pmpm_cost_t_12-9-6m_b4', 'rx_generic_pmpm_cost_t_6-3-0m_b4',
       'rx_overall_mbr_resp_pmpm_cost_t_6-3-0m_b4',
       'bh_ip_snf_mbr_resp_pmpm_cost_6to9m_b4',
       'rx_overall_dist_gpi6_pmpm_ct_t_6-3-0m_b4', 'auth_3mth_dc_no_ref',
       'auth_3mth_dc_snf', 'rx_phar_cat_humana_pmpm_ct_t_9-6-3m_b4',
       'bh_ip_snf_net_paid_pmpm_cost_0to3m_b4', 'auth_3mth_psychic',
       'auth_3mth_bh_acute', 'auth_3mth_acute_chf',
       'rx_overall_gpi_pmpm_ct_t_6-3-0m_b4', 'mcc_chf_pmpm_ct_t_9-6-3m_b4',
       'bh_urgent_care_copay_pmpm_cost_t_12-9-6m_b4', 'auth_3mth_acute_bld',
       'rx_gpi2_34_dist_gpi6_pmpm_ct', 'rx_maint_pmpm_cost_t_6-3-0m_b4',
```

```
        'cons_mobplus', 'lab_albumin_loinc_pmpm_ct',
        'rx_maint_net_paid_pmpm_cost_t_12-9-6m_b4',
        'rej_med_outpatient_visit_ct_pmpm_t_6-3-0m_b4',
        'rej_med_ip_snf_coins_pmpm_cost_t_9-6-3m_b4',
        'rx_gpi2_72_pmpm_ct_6to9m_b4',
        'med_physician_office_allowed_pmpm_cost_t_9-6-3m_b4',
        'auth_3mth_acute_res', 'auth_3mth_acute_dig',
        'auth_3mth_dc_acute_rehab', 'bh_ip_snf_mbr_resp_pmpm_cost_3to6m_b4',
        'total_physician_office_net_paid_pmpm_cost_t_9-6-3m_b4',
        'rx_branded_pmpm_ct_t_6-3-0m_b4',
        'med_outpatient_deduct_pmpm_cost_t_9-6-3m_b4', 'auth_3mth_non_er',
        'total_allowed_pmpm_cost_t_9-6-3m_b4', 'mabh_seg',]

for col in list3:
    train[col] = train[col].astype(str)

for col in list3:
    train[col] = LabelEncoder().fit_transform(train[col])
```

In [26]:

```
list4 = ['bh_ip_snf_mbr_resp_pmpm_cost_9to12m_b4', 'auth_3mth_post_acute_cer',
       'oontwk_mbr_resp_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_post_acute_mus',
       'hum_region', 'rx_nonmail_dist_gpi6_pmpm_ct_t_9-6-3m_b4',
       'bh_ip_snf_net_paid_pmpm_cost_6to9m_b4',
       'rej_med_er_net_paid_pmpm_cost_t_9-6-3m_b4',
       'med_outpatient_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'rx_nonbh_net_paid_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_post_acute_sns',
       'rx_gpi2_39_pmpm_cost_t_6-3-0m_b4',
       'total_ip_maternity_net_paid_pmpm_cost_t_12-9-6m_b4',
       'auth_3mth_acute_can', 'auth_3mth_post_acute', 'auth_3mth_facility',
       'rx_maint_pmpm_ct_t_6-3-0m_b4', 'auth_3mth_post_acute_men',
       'rx_mail_net_paid_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_home',
       'total_physician_office_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'auth_3mth_transplant', 'rev_cms_ansth_pmpm_ct',
       'rx_mail_mbr_resp_pmpm_cost_t_9-6-3m_b4',
       'med_outpatient_visit_ct_pmpm_t_12-9-6m_b4',
       'rx_nonbh_pmpm_ct_t_9-6-3m_b4', 'auth_3mth_acute',
       'auth_3mth_dc_left_ama', 'auth_3mth_acute_ccs_227',
       'auth_3mth_dc_custodial', 'total_med_net_paid_pmpm_cost_t_6-3-0m_b4',
       'rx_gpi2_90_dist_gpi6_pmpm_ct_9to12m_b4', 'sex_cd',
       'rx_gpi2_62_pmpm_cost_t_9-6-3m_b4',
       'rx_overall_gpi_pmpm_ct_t_12-9-6m_b4', 'auth_3mth_ltac', 'cons_hhcomp',
       'rx_nonmaint_dist_gpi6_pmpm_ct_t_12-9-6m_b4', 'auth_3mth_snf_post_hsp',
       'rx_nonotc_pmpm_cost_t_6-3-0m_b4', 'auth_3mth_acute_trm',
       'rej_total_physician_office_visit_ct_pmpm_0to3m_b4',
       'auth_3mth_snf_direct', 'auth_3mth_dc_home_health',
       'rx_gpi2_56_dist_gpi6_pmpm_ct_3to6m_b4', 'auth_3mth_acute_ner',
       'med_ambulance_coins_pmpm_cost_t_9-6-3m_b4', 'hedis_dia_hba1c_ge9',
       'ccsp_065_pmpm_ct', 'rx_maint_net_paid_pmpm_cost_t_9-6-3m_b4',
       'bh_ip_snf_admit_days_pmpm_t_9-6-3m_b4']

for col in list4:
    train[col] = train[col].astype(str)

for col in list4:
    train[col] = LabelEncoder().fit_transform(train[col])
```

In [27]:

```
categ_cols = train.select_dtypes(include='object')
categ_cols.columns
```

Out[27]:

```
Index([], dtype='object')
```

In [ ]:

```
train.info()
```

In [28]:

```
train.head()
```

Out[28]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pct_fmrkt_
0	0	66939	1	0.000000	7.910346	0.049413	0
1	1	225151	1	0.000000	1.730272	0.095624	60
2	2	696874	1	0.000000	5.015501	0.022398	30
3	3	117133	1	2.266667	4.049586	0.070407	50
4	4	22286	1	0.000000	0.618606	0.074862	88

5 rows × 366 columns

In [30]:
```
from sklearn.preprocessing import MinMaxScaler
df_std = MinMaxScaler().fit_transform(train)
df_new = pd.DataFrame(df_std, columns=train.columns)
df_new.head()
```

Out[30]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pct_fmrk
0	0.000000	0.068667	1.0	0.00000	0.253895	0.080931	
1	0.000001	0.230962	1.0	0.00000	0.055536	0.156618	
2	0.000002	0.714859	1.0	0.00000	0.160980	0.036685	
3	0.000003	0.120156	1.0	0.00059	0.129978	0.115316	
4	0.000004	0.022861	1.0	0.00000	0.019855	0.122613	

5 rows × 366 columns

In [31]:
```
test['covid_vaccination'] = LabelEncoder().fit_transform(test['covid_vaccination'])
test['covid_vaccination'].value_counts()
test
```

Out[31]:

	covid_vaccination
0	0
1	0
2	0
3	0
4	0
...	...
974837	0
974838	0
974839	0
974840	0
974841	0

974842 rows × 1 columns

In [32]:
```
data = pd.concat([df_new,test], axis=1)
data.head()
```

Out[32]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pct_fmrk
0	0.000000	0.068667	1.0	0.00000	0.253895	0.080931	
1	0.000001	0.230962	1.0	0.00000	0.055536	0.156618	
2	0.000002	0.714859	1.0	0.00000	0.160980	0.036685	
3	0.000003	0.120156	1.0	0.00059	0.129978	0.115316	
4	0.000004	0.022861	1.0	0.00000	0.019855	0.122613	

5 rows × 367 columns

In [33]:
```
# taking all records from minority group
minorityN = len(data[data.covid_vaccination == 1]) # get the total count of low-frequency group
minority_indices = data[data.covid_vaccination == 1].index
minority_sample = data.loc[minority_indices]
minority_sample
```

Out[33]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pc
11	0.000011	0.703111	1.0	0.000000	0.126735	0.073774	
14	0.000014	0.774533	1.0	0.000000	0.128916	0.157680	
17	0.000017	0.946569	1.0	0.000000	0.037677	0.017380	
19	0.000019	0.108334	1.0	0.000000	0.128178	0.163629	
23	0.000024	0.502195	1.0	0.000000	0.231772	0.158584	

...	...	...	...	...	...	...
974808	0.999966	0.668001	1.0	0.006843	0.313572	0.126572
974817	0.999975	0.333033	1.0	0.000000	0.237097	0.145730
974822	0.999981	0.253427	1.0	0.000000	0.232773	0.143694
974823	0.999982	0.915395	1.0	0.001429	0.194265	0.097758
974833	0.999992	0.612678	1.0	0.003232	0.100701	0.142432

169453 rows × 367 columns

In [34]:
```
# Perform undersampling majority group
majority_indices = data[data.covid_vaccination == 0].index
random_indices = np.random.choice(majority_indices, minorityN, replace=False) # use the low-frequency group count
majority_sample = data.loc[random_indices]
majority_sample
```

Out[34]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pc
17776	0.018235	0.033843	1.0	0.0	0.164473	0.161627	
509904	0.523064	0.125096	1.0	0.0	0.096248	0.095483	
72126	0.073987	0.759735	1.0	0.0	0.096693	0.212590	
430426	0.441535	0.138025	1.0	0.0	0.096288	0.150780	
776291	0.796326	0.513580	1.0	0.0	0.137545	0.109746	
...	...	...	...	...	...	...	
612594	0.628404	0.302557	1.0	0.0	0.406595	0.000000	
953191	0.977791	0.843292	1.0	0.0	0.051394	0.452325	
667632	0.684862	0.868530	1.0	0.0	0.141712	0.185894	
488686	0.501298	0.175757	1.0	0.0	0.070279	0.045900	
369471	0.379006	0.254813	1.0	0.0	0.252408	0.200909	

169453 rows × 367 columns

In [35]:
```
merged_sample = pd.concat([minority_sample, majority_sample], ignore_index=True) # merging all the low-frequency
merged_sample.to_csv("preprocessed_data.csv")
merged_sample
```

Out[35]:

	Unnamed: 0	ID	auth_3mth_post_acute_dia	rx_gpi2_72_pmpm_cost_6to9m_b4	atlas_pct_laccess_child15	atlas_recfacpth14	atlas_pc
0	0.000011	0.703111	1.0	0.0	0.126735	0.073774	
1	0.000014	0.774533	1.0	0.0	0.128916	0.157680	
2	0.000017	0.946569	1.0	0.0	0.037677	0.017380	
3	0.000019	0.108334	1.0	0.0	0.128178	0.163629	
4	0.000024	0.502195	1.0	0.0	0.231772	0.158584	
...	...	...	...	...	...	...	
338901	0.628404	0.302557	1.0	0.0	0.406595	0.000000	
338902	0.977791	0.843292	1.0	0.0	0.051394	0.452325	
338903	0.684862	0.868530	1.0	0.0	0.141712	0.185894	
338904	0.501298	0.175757	1.0	0.0	0.070279	0.045900	
338905	0.379006	0.254813	1.0	0.0	0.252408	0.200909	

338906 rows × 367 columns

In [ ]:
```
```