```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")

df=pd.read_csv("h1n1_vaccine_prediction.csv")

df.head()
```

```
   unique_id  h1n1_worry  h1n1_awareness  antiviral_medication  \
0          0         1.0             0.0                   0.0
1          1         3.0             2.0                   0.0
2          2         1.0             1.0                   0.0
3          3         1.0             1.0                   0.0
4          4         2.0             1.0                   0.0

   contact_avoidance  bought_face_mask  wash_hands_frequently  \
0                0.0               0.0                    0.0
1                1.0               0.0                    1.0
2                1.0               0.0                    0.0
3                1.0               0.0                    1.0
4                1.0               0.0                    1.0

   avoid_large_gatherings  reduced_outside_home_cont  avoid_touch_face
...  \
0                     0.0                        1.0               1.0
...
1                     0.0                        1.0               1.0
...
2                     0.0                        0.0               0.0
...
3                     1.0                        0.0               0.0
...
4                     1.0                        0.0               1.0
...

    race     sex               income_level  marital_status
housing_status  \
0  White  Female             Below Poverty     Not Married
Own
1  White    Male             Below Poverty     Not Married
Rent
2  White    Male  <= $75,000, Above Poverty     Not Married
Own
3  White  Female             Below Poverty     Not Married
Rent
4  White  Female  <= $75,000, Above Poverty         Married
Own
```

```
           employment                census_msa  no_of_adults
no_of_children  \
0  Not in Labor Force                   Non-MSA           0.0
0.0
1            Employed  MSA, Not Principle  City           0.0
0.0
2            Employed  MSA, Not Principle  City           2.0
0.0
3  Not in Labor Force       MSA, Principle City           0.0
0.0
4            Employed  MSA, Not Principle  City           1.0
0.0

   h1n1_vaccine
0             0
1             0
2             0
3             0
4             0

[5 rows x 34 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26707 entries, 0 to 26706
Data columns (total 34 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   unique_id                    26707 non-null  int64
 1   h1n1_worry                   26615 non-null  float64
 2   h1n1_awareness               26591 non-null  float64
 3   antiviral_medication         26636 non-null  float64
 4   contact_avoidance            26499 non-null  float64
 5   bought_face_mask             26688 non-null  float64
 6   wash_hands_frequently        26665 non-null  float64
 7   avoid_large_gatherings       26620 non-null  float64
 8   reduced_outside_home_cont    26625 non-null  float64
 9   avoid_touch_face             26579 non-null  float64
 10  dr_recc_h1n1_vacc            24547 non-null  float64
 11  dr_recc_seasonal_vacc        24547 non-null  float64
 12  chronic_medic_condition      25736 non-null  float64
 13  cont_child_undr_6_mnths      25887 non-null  float64
 14  is_health_worker             25903 non-null  float64
 15  has_health_insur             14433 non-null  float64
 16  is_h1n1_vacc_effective       26316 non-null  float64
 17  is_h1n1_risky                26319 non-null  float64
 18  sick_from_h1n1_vacc          26312 non-null  float64
 19  is_seas_vacc_effective       26245 non-null  float64
```

```
 20  is_seas_risky              26193 non-null  float64
 21  sick_from_seas_vacc        26170 non-null  float64
 22  age_bracket                26707 non-null  object
 23  qualification              25300 non-null  object
 24  race                       26707 non-null  object
 25  sex                        26707 non-null  object
 26  income_level               22284 non-null  object
 27  marital_status             25299 non-null  object
 28  housing_status             24665 non-null  object
 29  employment                 25244 non-null  object
 30  census_msa                 26707 non-null  object
 31  no_of_adults               26458 non-null  float64
 32  no_of_children             26458 non-null  float64
 33  h1n1_vaccine               26707 non-null  int64
dtypes: float64(23), int64(2), object(9)
memory usage: 6.9+ MB
```

```
df.isnull().sum()/len(df)*100
```

```
unique_id                     0.000000
h1n1_worry                    0.344479
h1n1_awareness                0.434343
antiviral_medication          0.265848
contact_avoidance             0.778822
bought_face_mask              0.071142
wash_hands_frequently         0.157262
avoid_large_gatherings        0.325757
reduced_outside_home_cont     0.307036
avoid_touch_face              0.479275
dr_recc_h1n1_vacc             8.087767
dr_recc_seasonal_vacc         8.087767
chronic_medic_condition       3.635751
cont_child_undr_6_mnths       3.070356
is_health_worker              3.010447
has_health_insur             45.957989
is_h1n1_vacc_effective        1.464036
is_h1n1_risky                 1.452803
sick_from_h1n1_vacc           1.479013
is_seas_vacc_effective        1.729884
is_seas_risky                 1.924589
sick_from_seas_vacc           2.010709
age_bracket                   0.000000
qualification                 5.268282
race                          0.000000
sex                           0.000000
income_level                 16.561201
marital_status                5.272026
housing_status                7.645936
employment                    5.477965
census_msa                    0.000000
no_of_adults                  0.932340
```

```
no_of_children                0.932340
h1n1_vaccine                  0.000000
dtype: float64
```

```
df.drop('unique_id',axis=1,inplace=True)
```

```
df['dr_recc_h1n1_vacc']=df['dr_recc_h1n1_vacc'].fillna(df['dr_recc_h1n
1_vacc'].mode()[0])
```

```
df['dr_recc_seasonal_vacc']=df['dr_recc_seasonal_vacc'].fillna(df['dr_
recc_seasonal_vacc'].mode()[0])
```

```
df['has_health_insur']=df['has_health_insur'].fillna(df['has_health_in
sur'].mode()[0])
```

```
df['qualification']=df['qualification'].fillna(df['qualification'].mod
e()[0])
```

```
df['income_level']=df['income_level'].fillna(df['income_level'].mode()
[0])
```

```
df['marital_status']=df['marital_status'].fillna(df['marital_status'].
mode()[0])
```

```
df['housing_status']=df['housing_status'].fillna(df['housing_status'].
mode()[0])
```

```
df['employment']=df['employment'].fillna(df['employment'].mode()[0])
```

```
df.dropna(inplace=True)
```

```
df.reset_index(drop=True,inplace=True)
```

```
df.shape
```

```
(24803, 33)
```

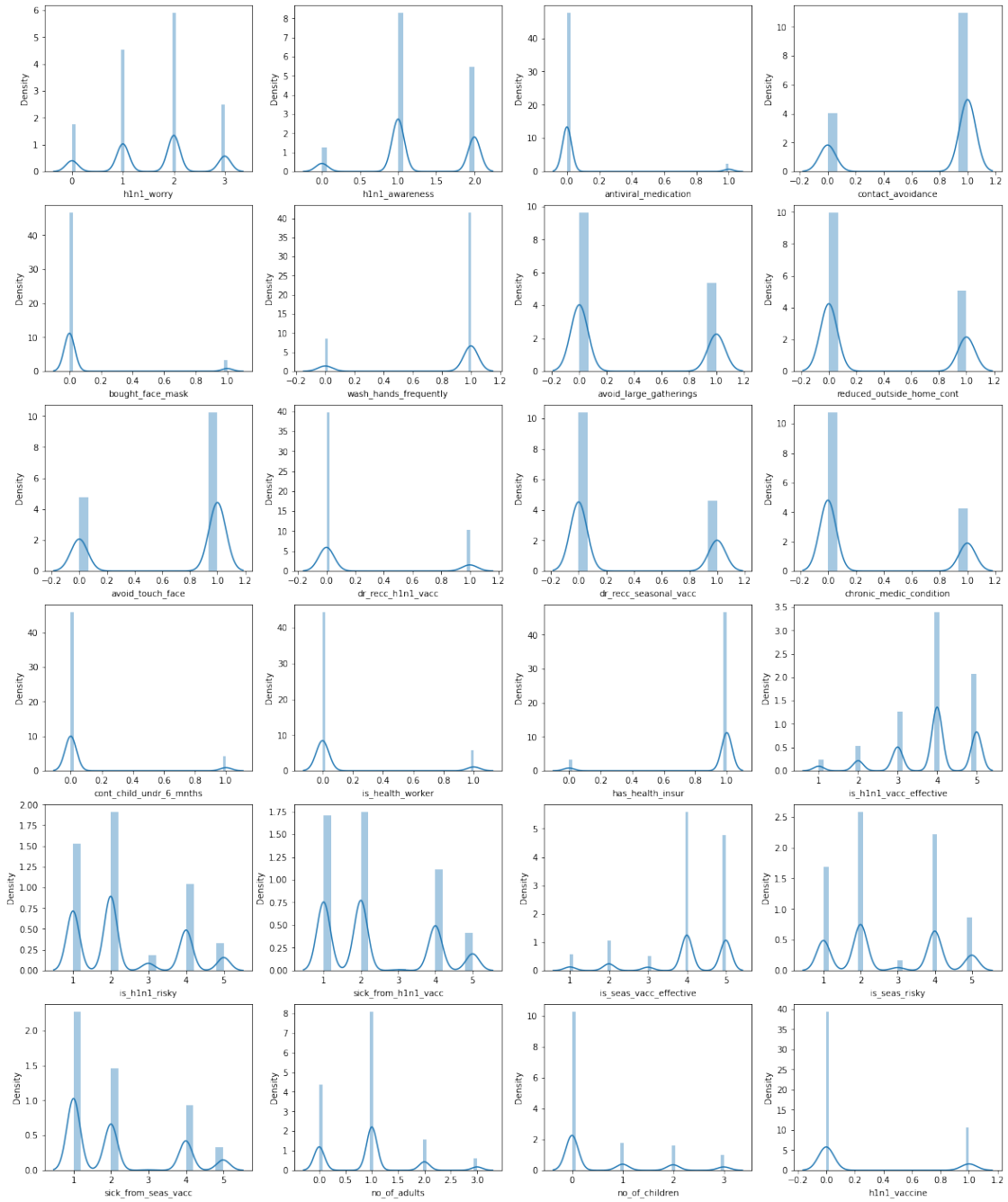
```
cat_cols=df.select_dtypes(include="O").columns
num_cols=df.select_dtypes(include=["int","float"]).columns
```

```
for i in cat_cols:
    print(i,":",df[i].unique())
```

```
age_bracket : ['55 - 64 Years' '35 - 44 Years' '18 - 34 Years' '65+
Years'
 '45 - 54 Years']
qualification : ['< 12 Years' '12 Years' 'College Graduate' 'Some
College']
race : ['White' 'Black' 'Other or Multiple' 'Hispanic']
sex : ['Female' 'Male']
income_level : ['Below Poverty' '<= $75,000, Above Poverty' '>
$75,000']
marital_status : ['Not Married' 'Married']
housing_status : ['Own' 'Rent']
```

```
employment : ['Not in Labor Force' 'Employed' 'Unemployed']
census_msa : ['Non-MSA' 'MSA, Not Principle  City' 'MSA, Principle
City']
```

```
plt.figure(figsize=(20,25))
count=1
for i in num_cols:
    plt.subplot(6,4,count)
    sns.distplot(df[i])
    count+=1
```



```
df.describe()
```

```
        h1n1_worry  h1n1_awareness  antiviral_medication
contact_avoidance  \
count  24803.00000    24803.000000          24803.000000
24803.000000
mean       1.62315        1.279160              0.048099
0.731121
std        0.90275        0.608288              0.213980
0.443386
min        0.00000        0.000000              0.000000
0.000000
25%        1.00000        1.000000              0.000000
0.000000
50%        2.00000        1.000000              0.000000
1.000000
75%        2.00000        2.000000              0.000000
1.000000
max        3.00000        2.000000              1.000000
1.000000

       bought_face_mask  wash_hands_frequently  avoid_large_gatherings
\
count      24803.000000           24803.000000            24803.000000

mean           0.067895               0.828690                0.358586

std            0.251571               0.376787                0.479595

min            0.000000               0.000000                0.000000

25%            0.000000               1.000000                0.000000

50%            0.000000               1.000000                0.000000

75%            0.000000               1.000000                1.000000

max            1.000000               1.000000                1.000000


       reduced_outside_home_cont  avoid_touch_face  dr_recc_h1n1_vacc
...  \
count               24803.000000      24803.000000       24803.000000
...
mean                    0.336290          0.682135           0.205862
...
std                     0.472449          0.465656           0.404338
...
min                     0.000000          0.000000           0.000000
```

```
...
25%                    0.000000          0.000000           0.000000
...
50%                    0.000000          1.000000           0.000000
...
75%                    1.000000          1.000000           0.000000
...
max                    1.000000          1.000000           1.000000
...

       has_health_insur  is_h1n1_vacc_effective  is_h1n1_risky  \
count      24803.000000            24803.000000    24803.00000
mean           0.932831                3.868605        2.34774
std            0.250320                1.001426        1.28707
min            0.000000                1.000000        1.00000
25%            1.000000                3.000000        1.00000
50%            1.000000                4.000000        2.00000
75%            1.000000                5.000000        4.00000
max            1.000000                5.000000        5.00000

       sick_from_h1n1_vacc  is_seas_vacc_effective  is_seas_risky  \
count         24803.000000            24803.000000    24803.00000
mean              2.357981                4.037254        2.73096
std               1.362143                1.078777        1.38686
min               1.000000                1.000000        1.00000
25%               1.000000                4.000000        2.00000
50%               2.000000                4.000000        2.00000
75%               4.000000                5.000000        4.00000
max               5.000000                5.000000        5.00000

       sick_from_seas_vacc  no_of_adults  no_of_children  h1n1_vaccine

count         24803.000000  24803.000000    24803.000000  24803.000000

mean              2.115671      0.894610        0.542959      0.214087

std               1.331195      0.752345        0.933240      0.410196

min               1.000000      0.000000        0.000000      0.000000

25%               1.000000      0.000000        0.000000      0.000000

50%               2.000000      1.000000        0.000000      0.000000

75%               4.000000      1.000000        1.000000      0.000000

max               5.000000      3.000000        3.000000      1.000000
```

```
[8 rows x 24 columns]
```

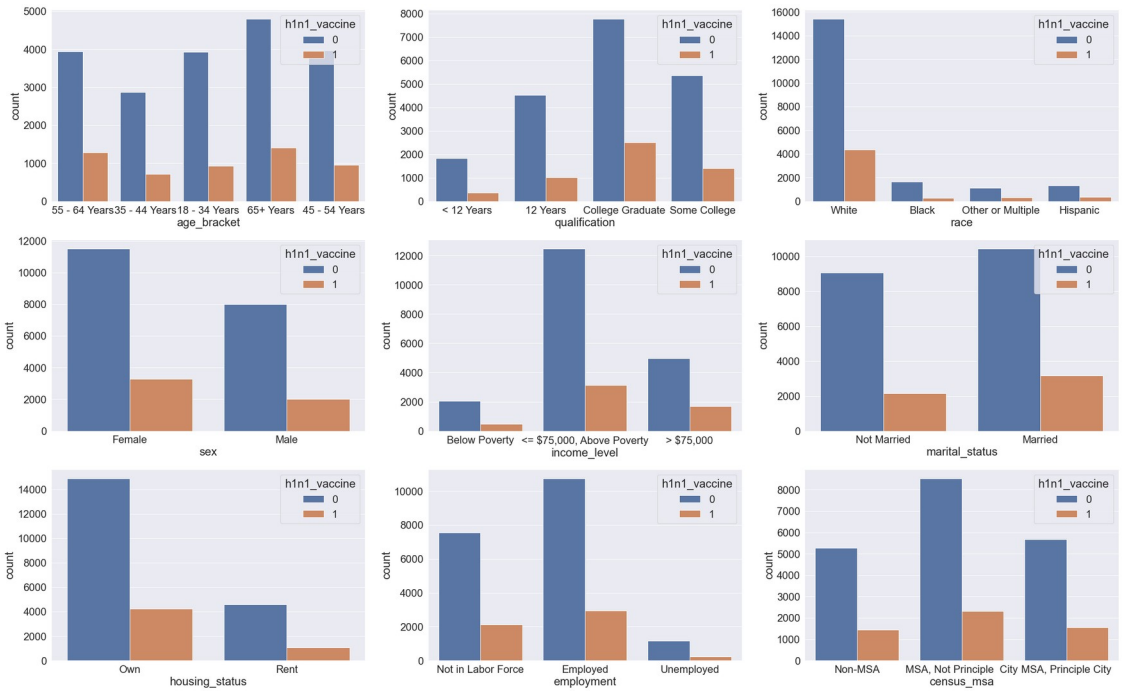
```
cat_cols
```

```
Index(['age_bracket', 'qualification', 'race', 'sex', 'income_level',
       'marital_status', 'housing_status', 'employment',
'census_msa'],
      dtype='object')
```

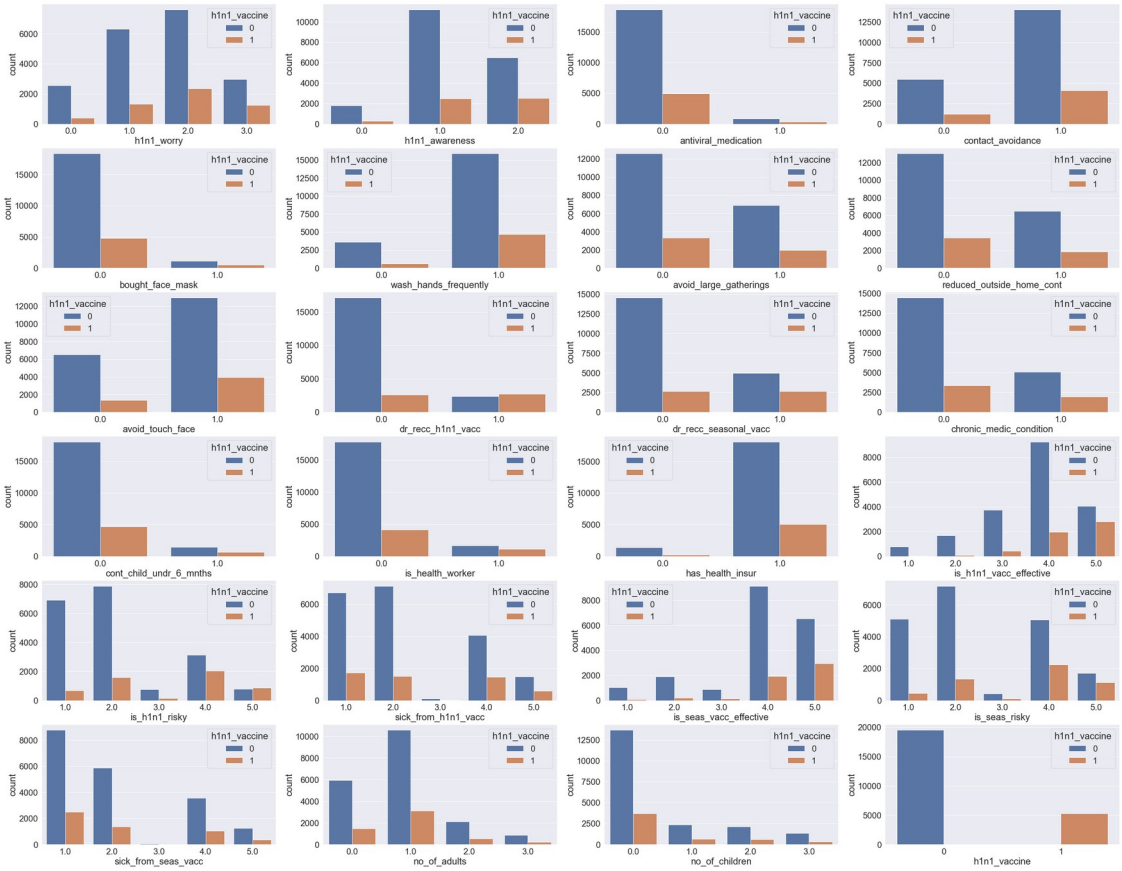
```
num_cols
```

```
Index(['h1n1_worry', 'h1n1_awareness', 'antiviral_medication',
       'contact_avoidance', 'bought_face_mask',
'wash_hands_frequently',
       'avoid_large_gatherings', 'reduced_outside_home_cont',
       'avoid_touch_face', 'dr_recc_h1n1_vacc',
'dr_recc_seasonal_vacc',
       'chronic_medic_condition', 'cont_child_undr_6_mnths',
       'is_health_worker', 'has_health_insur',
'is_h1n1_vacc_effective',
       'is_h1n1_risky', 'sick_from_h1n1_vacc',
'is_seas_vacc_effective',
       'is_seas_risky', 'sick_from_seas_vacc', 'no_of_adults',
       'no_of_children', 'h1n1_vaccine'],
      dtype='object')
```

```
sns.set(font_scale=2)
plt.figure(figsize=(40,25))
count=1
for i in cat_cols:
    plt.subplot(3,3,count)
    sns.countplot(df[i],hue=df['h1n1_vaccine'])
    count+=1
```




```
plt.figure(figsize=(50,40))
count=1
for i in num_cols:
    plt.subplot(6,4,count)
    sns.countplot(df[i],hue=df['h1n1_vaccine'])
    count+=1
```




```
from sklearn.preprocessing import LabelEncoder

le=LabelEncoder()

for i in cat_cols:
    df[i]=le.fit_transform(df[i])

x=df.iloc[:,:-1]
y=df.iloc[:,-1]

from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,rand
om_state=123)

from sklearn.preprocessing import StandardScaler

sc=StandardScaler()

x_train=sc.fit_transform(x_train)
x_test=sc.fit_transform(x_test)

from sklearn.linear_model import LogisticRegression

reg=LogisticRegression()
```

```
reg.fit(x_train,y_train)
```

```
LogisticRegression()
```

```
y_pred_train=reg.predict(x_train)
y_pred_test=reg.predict(x_test)
```

```
from sklearn.metrics import accuracy_score
```

```
print(accuracy_score(y_train,y_pred_train))
```

```
0.8353492591472634
```

```
print(accuracy_score(y_test,y_pred_test))
```

```
0.8385406168111268
```