Substance Abuse Treatment Analysis

By Hakeem Lawrence

Dataset presented by Enterprise DNA

The Brief

It is your job as an analyst to prepare an analysis report about the Substance Abuse dataset. The following questions must be addressed in your report.

1. Compare different hospitalization programs. What conclusion(s) can you draw from it?
2. What are key drivers of different types of primary mental health diagnosis?
3. Demographic analysis about different types of primary mental health diagnosis?
4. What other analysis would you like to have?
5. What other recommendations would you like to make?

Dataset Variables

Dataset Variables\ Admission Date: The date at which the client entered treatment\ PPID: Unique client ID\ program: Type of recovery program used\ Age: Age of client\ Gender: Sex category specified by client\ RaceEthinicty: Racial group specified by client\ MHDx: Mental Health Disorder Treatment\ SUDx: Substance Use Disorder Treatment\ Medx: Medical Disorder Treatment\ PsychAdmit: Current Pysch patient\ DLA1: First daily living assessment\ DLA2: 2nd daily living assessment

Importing libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Data Inspection

In [2]:
```
substanceAbuse = pd.read_excel('Substance Abuse.xlsx')
substanceAbuse.head()
```

Out[2]:

	Admission Date	PPID	Program	Age	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	DLA1
0	2022-01-13	A234282	Intervention	34	F	Other	Depression	Alcohol	2	1	3.69
1	2022-02-18	A232412	Intervention	26	M	NonHispWhite	Trauma	Opioid	0	0	4.22
2	2022-01-28	A259052	Intervention	62	M	NativeAm	Depression	Opioid	0	1	4.17
3	2022-01-30	A353421	Intervention	34	F	NonHispWhite	Depression	Alcohol	0	0	4.11
4	2022-03-28	A302351	UsualCare	46	M	NonHispBlack	Trauma	Opioid	0	1	4.19

In [3]:
```
# creating duplicate dataframe to preserve the original

substanceAbuse2 = substanceAbuse
```

In [4]:
```
substanceAbuse2.columns
```

Out[4]:
```
Index(['Admission Date', 'PPID', 'Program', 'Age', 'Gender', 'RaceEthnicity',
       'MHDx', 'SUDx', 'MedDx', 'PsychAdmit', 'DLA1', 'DLA2'],
      dtype='object')
```

In [5]:
```
substanceAbuse2.shape
```

Out[5]:
```
(479, 12)
```

In [6]:
```
substanceAbuse2.isna().sum().sum()
```

Out[6]:
```
0
```

In [7]:
```
substanceAbuse2.duplicated().sum()
```

Out[7]:
```
0
```

This dataset is relatively clean. It has no duplicate observations or missing values. The next thing we will inspect is the datatypes for each variable. We will need to confirm that their is consistency among value labels.

In [8]:
```
substanceAbuse2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 479 entries, 0 to 478
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Admission Date  479 non-null    datetime64[ns]
 1   PPID            479 non-null    object
 2   Program         479 non-null    object
 3   Age             479 non-null    int64
 4   Gender          479 non-null    object
```

```
 5   RaceEthnicity   479 non-null    object
 6   MHDx            479 non-null    object
 7   SUDx            479 non-null    object
 8   MedDx           479 non-null    object
 9   PsychAdmit      479 non-null    int64
 10  DLA1            479 non-null    float64
 11  DLA2            479 non-null    float64
dtypes: datetime64[ns](1), float64(2), int64(2), object(7)
memory usage: 45.0+ KB
```

In [9]:
```
# we do not need the client ID column so we will drop it

substanceAbuse2.drop('PPID', axis=1, inplace=True)
```

In [10]:
```
substanceAbuse2.head()
```

Out[10]:

	Admission Date	Program	Age	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	DLA1	DLA2
0	2022-01-13	Intervention	34	F	Other	Depression	Alcohol	2	1	3.69	4.13
1	2022-02-18	Intervention	26	M	NonHispWhite	Trauma	Opioid	0	0	4.22	4.68
2	2022-01-28	Intervention	62	M	NativeAm	Depression	Opioid	0	1	4.17	4.78
3	2022-01-30	Intervention	34	F	NonHispWhite	Depression	Alcohol	0	0	4.11	4.46
4	2022-03-28	UsualCare	46	M	NonHispBlack	Trauma	Opioid	0	1	4.19	4.25

We need to inspect the MedDx column since it contains numbers but has a object data type.

In [11]:
```
# we need to remove the '+' sign from 3 in order to make it the correct data type

substanceAbuse2['MedDx'].value_counts()
```

Out[11]:
```
0      200
1      157
2       95
3+      27
Name: MedDx, dtype: int64
```

In [12]:
```
substanceAbuse2['MedDx'].replace('3+',3,inplace=True)
```

In [13]:
```
substanceAbuse2['MedDx'].value_counts()
```

Out[13]:
```
0     200
1     157
2      95
3      27
Name: MedDx, dtype: int64
```

We cleaned the suspicous column. The next thing we will do is create an Age Group variable. This will help segment the clients by age groups.

In [14]:
```
# feature engineering an Age Group variable

substanceAbuse2['Age Group'] = substanceAbuse2['Age']

substanceAbuse2['Age Group']
```

Out[14]:
```
0      34
1      26
2      62
3      34
4      46
       ..
474    58
475    47
476    39
477    63
478    66
Name: Age Group, Length: 479, dtype: int64
```

In [15]:
```
# Assigning labels using Numpy Vectorization

conditions =[
    substanceAbuse2['Age'] < 30,
    ((substanceAbuse2['Age'] >= 30) & (substanceAbuse2['Age'] <40)),
    ((substanceAbuse2['Age'] >= 40) & (substanceAbuse2['Age'] <50)),
    ((substanceAbuse2['Age'] >= 50) & (substanceAbuse2['Age'] <60)),
    ((substanceAbuse2['Age'] >= 60) & (substanceAbuse2['Age'] <70)),
    substanceAbuse2['Age'] >= 70
]


groups = [
    'Under 30',
    "30's",
    "40's",
    "50's",
    "60's",
    "70's and up"
]


substanceAbuse2['Age Group']= np.select(conditions, groups, default= 'NA')
substanceAbuse2['Age Group'].value_counts()
```

Out[15]:
```
40's           133
50's           121
30's           106
60's            54
Under 30        51
70's and up     14
Name: Age Group, dtype: int64
```

In [16]:
```
substanceAbuse2['Age Group'].value_counts(normalize=True)
```

Out[16]:
```
40's           0.277662
50's           0.252610
30's           0.221294
60's           0.112735
Under 30       0.106472
70's and up    0.029228
Name: Age Group, dtype: float64
```

Most of the ages in this dataset ranges from 30 to 50. Possible

Next we will seperate the data into discrete and continous variables. We'll use frequencies and proportions for discrete variables and descriptive stats for the continous variables.

In [17]:
```
continuousVariables = ['Age', 'DLA1', 'DLA2']
```

In [18]:
```
discreteVariables = ['Program', 'Gender', 'RaceEthnicity','MHDx', 'SUDx', 'MedDx', 'Psyc
```

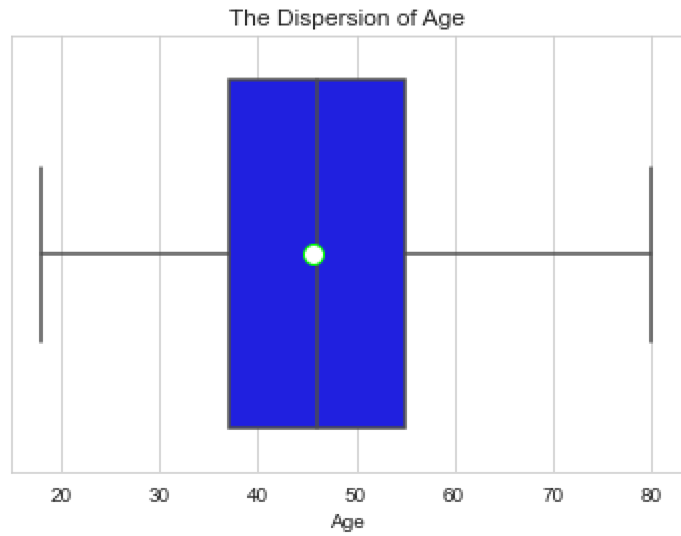
Descriptive Statistics

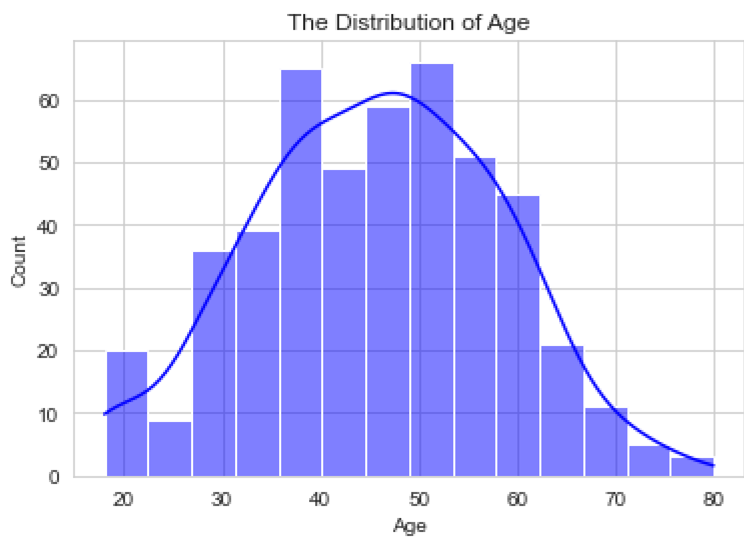
In [19]:
```
# Descriptive Stats on ratio numeric variables. PyschAdmit will be grouped as discrete.

substanceAbuse2[continuousVariables].describe()
```

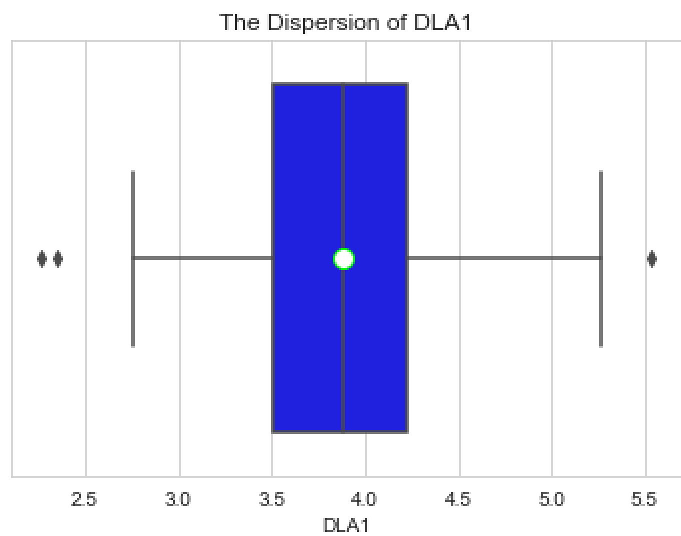
Out[19]:

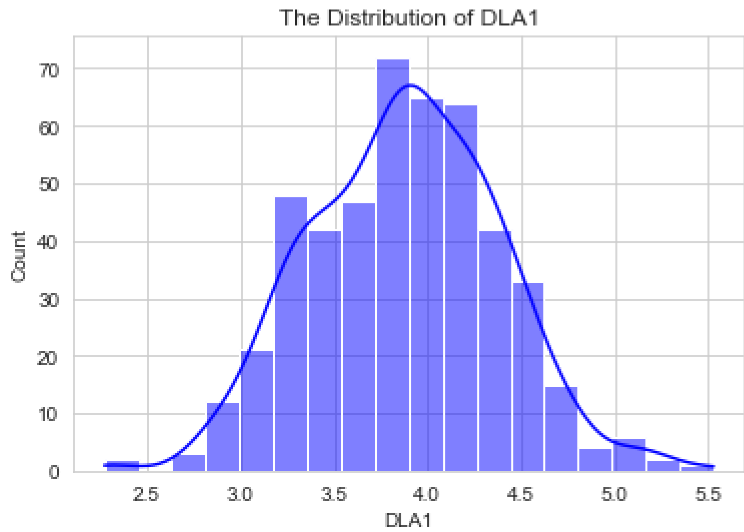
	Age	DLA1	DLA2
count	479.000000	479.000000	479.000000
mean	45.661795	3.878789	4.053236
std	12.485909	0.501458	0.587929
min	18.000000	2.270000	2.290000
25%	37.000000	3.510000	3.650000
50%	46.000000	3.880000	4.040000
75%	55.000000	4.230000	4.460000
max	80.000000	5.530000	6.020000

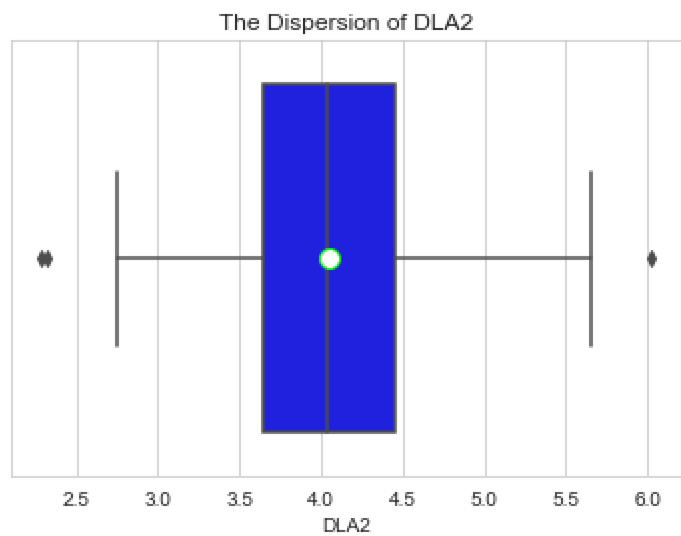
In [20]:
```
# Displaying box-plots and histograms for each continous variable to visualize the sprea

sns.set_style('whitegrid')
def boxPlotAndHistPlot(var):
    'This function creates boxplots and histplots for all contious variables '
    plt.figure(figsize= (6,4))
    sns.boxplot(x =var, data=  substanceAbuse2,\
                color = 'blue', showmeans= True,\
               meanprops={'marker': 'o',
                          'markerfacecolor':'white',
                          'markeredgecolor': 'lime',
                          'markersize': '10'})
    plt.title(f'The Dispersion of {var}')
    plt.show()
    sns.histplot(x=var, data= substanceAbuse2, color='blue', kde=True)
    plt.title(f'The Distribution of {var}')
    plt.show()



for i in continuousVariables:
    boxPlotAndHistPlot(i)
```

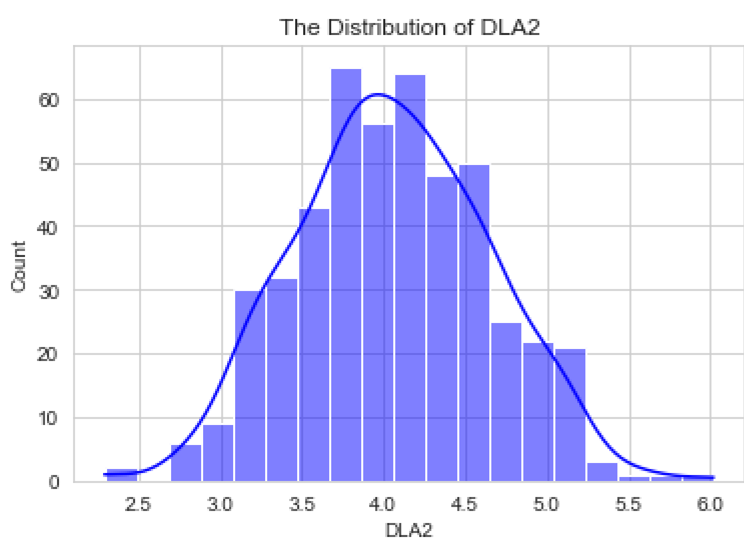

The Dispersion of Age
20
30
40
50
60
70
80
Age


The Distribution of Age
Count
0
10
20
30
40
50
60
20
30
40
50
60
70
80
Age


The Dispersion of DLA1
2.5
3.0
3.5
4.0
4.5
5.0
5.5
DLA1







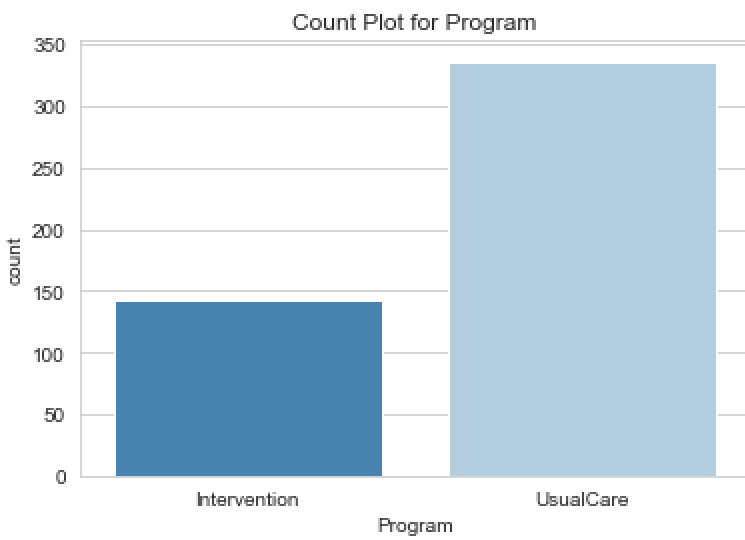
In [21]:
```
DLA1IQR = substanceAbuse2['DLA1'].quantile(.75) - substanceAbuse2['DLA1'].quantile(.25)
DLA2IQR = substanceAbuse2['DLA2'].quantile(.75) - substanceAbuse2['DLA2'].quantile(.25)

print(f'The IQR for DLA1 is: {DLA1IQR.round(2)}')
print(f'The IQR for DLA2 is: {DLA2IQR.round(2)}')
```

```
The IQR for DLA1 is: 0.72
The IQR for DLA2 is: 0.81
```

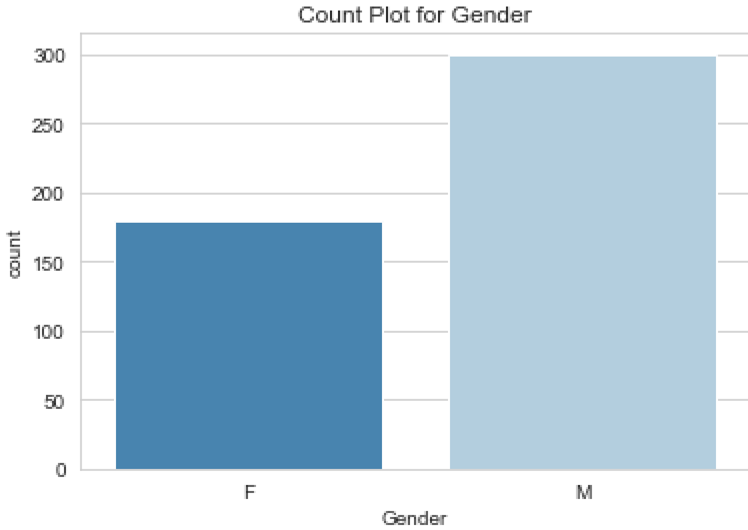
Our boxplots and histograms shows that all continious variables have a normal distribution.\ Age: This variable doesn't have any outliers. Most of the patients are around 37-55 years old or +/- 1 std.dev from the mean. \ DLA1 & DLA2: Both have a few outliers with DLA1 showing less dispersion than DLA2.This suggests that DLA2's results vary more than DLA1's results.

Now we will view the frequencies and proportions of our discrete variables

In [53]:
```
# viewing frequencies and proportions for discrete variables

def countPlot(var):
    'This function creates bar plots and Normalized distribution tables for all discrete
    plt.figure(figsize=(6,4))
    sns.countplot(x=var, data=substanceAbuse2, palette = 'Blues_r')
    plt.title(f'Count Plot for {var}')
    plt.show()

    proportions = substanceAbuse2[var].value_counts(normalize=True).round(3).to_frame().

    print(' ')
    print(f'Normalized Proportions of {var}')
    print(' ')
    print(proportions)
    print('-'*30)

for i in discreteVariables:
    countPlot(i)
```

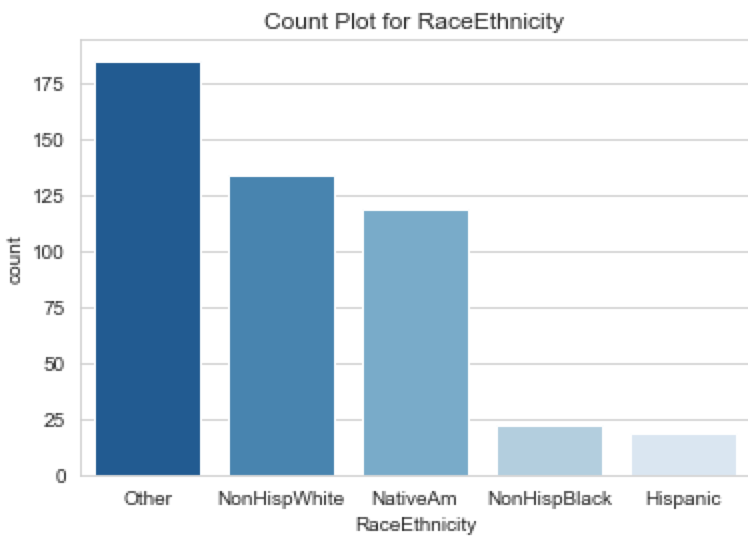



```
Normalized Proportions of Program

        Program  Normalized
0     UsualCare       0.701
1  Intervention       0.299
------------------------------
```



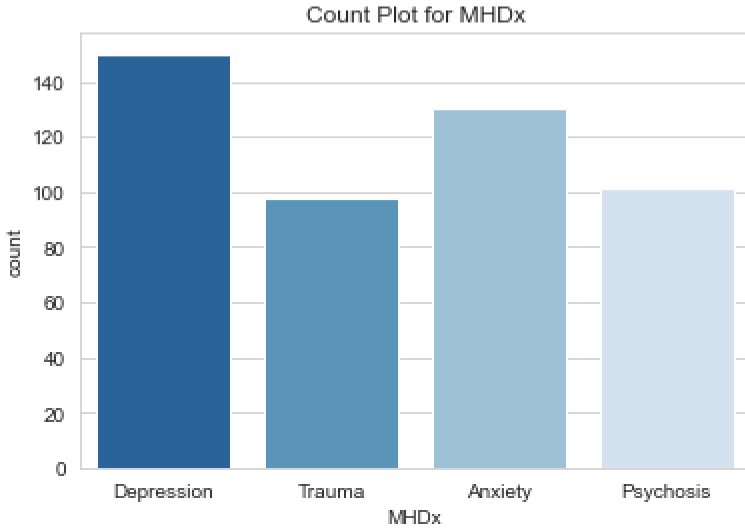
Normalized Proportions of Gender

	Gender	Normalized
0	M	0.626
1	F	0.374

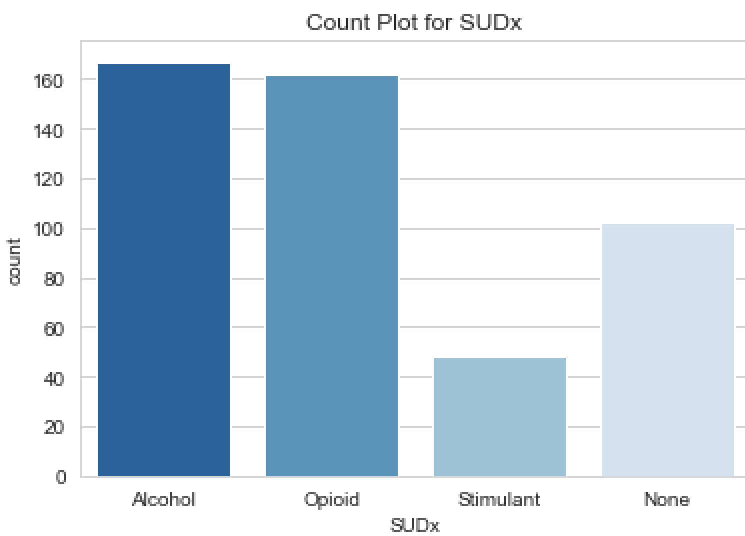


Normalized Proportions of RaceEthnicity

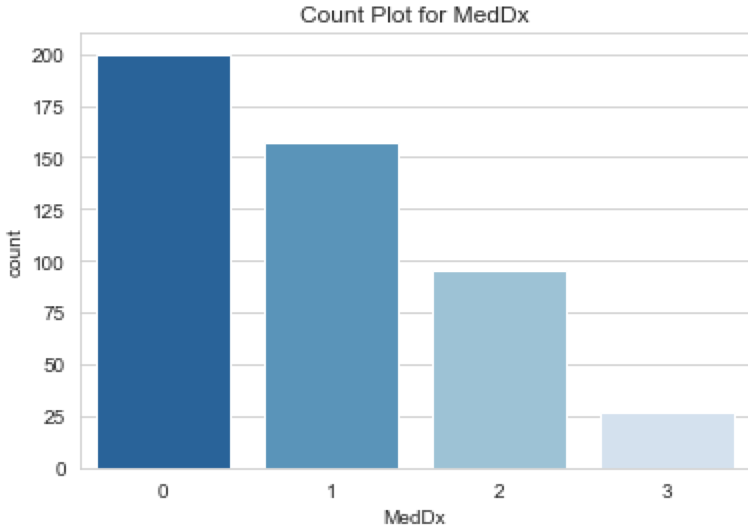
	RaceEthnicity	Normalized
0	Other	0.386
1	NonHispWhite	0.280
2	NativeAm	0.248
3	NonHispBlack	0.046
4	Hispanic	0.040



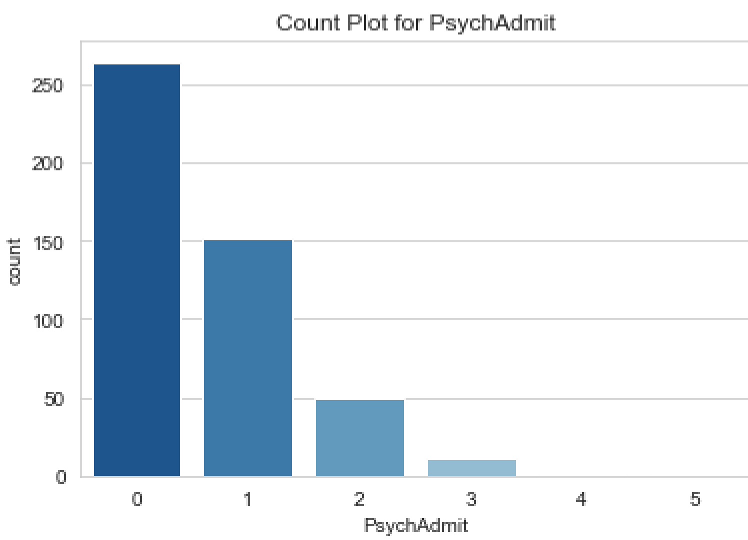

```
Normalized Proportions of MHDx

         MHDx  Normalized
0  Depression       0.313
1     Anxiety       0.271
2   Psychosis       0.211
3      Trauma       0.205
------------------------------
```



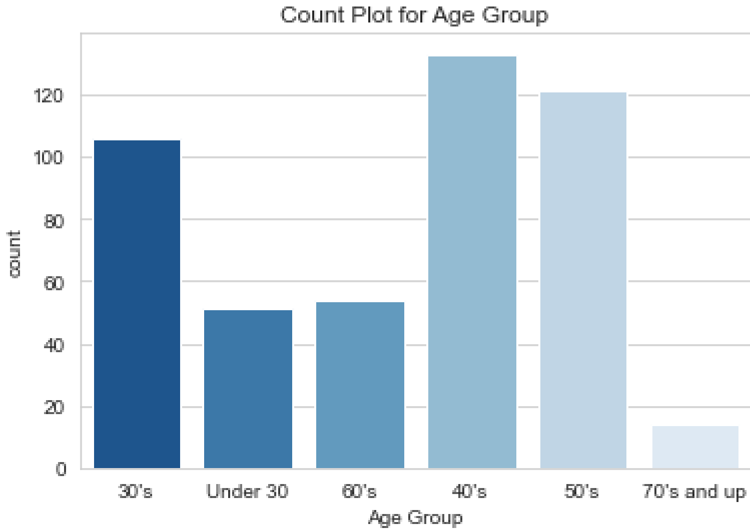

```
Normalized Proportions of SUDx

        SUDx  Normalized
0    Alcohol       0.349
1     Opioid       0.338
2       None       0.213
3  Stimulant       0.100
------------------------------
```




```
Normalized Proportions of MedDx

   MedDx  Normalized
0      0       0.418
1      1       0.328
2      2       0.198
3      3       0.056
------------------------------
```




```
Normalized Proportions of PsychAdmit

   PsychAdmit  Normalized
0           0       0.551
1           1       0.317
2           2       0.104
3           3       0.023
4           5       0.002
5           4       0.002
------------------------------
```




```
Normalized Proportions of Age Group

     Age Group  Normalized
0         40's       0.278
1         50's       0.253
2         30's       0.221
3         60's       0.113
4     Under 30       0.106
5  70's and up       0.029
------------------------------
```

The value counts have validated that there are no erroneous words or characters in each categorical column. Usual Care has more than twice the observations than the intervention program and accounts for 70% of data. Gender is highly unbalanced with 62% of patients reporting to be male. Nearly 40% of patient's ethnicities have been grouped into the 'other' category while Whites make up nearly 30% and Native Americans 25%. Other groups of people who may not be accounted for are Asian Indian, Polish, Irish or Italian. Depression accounts for nearly 1/3 of mental health diagnosis while anxiety follows. Alcohol and Opioid addictions accounts for 68% of all substances used. Most patients have less than 2 medical diagnosis and majority have been admitted to Psych less than 2 times. In fact, 86% of patients were admitted to Pysch at most one time. We can consider all Psych visits above one to be abnormal.

Addressing Analytics Requests

1. Compare different hospitalization programs. What conclusion(s) can you draw from it?

In [23]:
```
dla1Comparison = substanceAbuse2.groupby('Program').agg({'DLA1':['mean', 'std']})      #c
dlalevel0 = dla1Comparison.columns.get_level_values(0)     # get labels from first level
dlalevel1 = dla1Comparison.columns.get_level_values(1)     # get lables from second level
dla1Comparison.columns = dlalevel0 + '_' + dlalevel1          #  Replace old column names
dla1Comparison.reset_index().head()                        # Reset the index to get them on t
```

Out[23]:

	Program	DLA1_mean	DLA1_std
0	Intervention	3.908741	0.479951
1	UsualCare	3.866042	0.510502

There is barely any difference between the means in DLA1 scores for treatment programs. However, ususal care has a slightly greater standard deviation than intervention.

In [24]:
```
dla2Comparison =substanceAbuse2.groupby('Program').agg({'DLA2': ['mean','std']})
dla2level0 = dla2Comparison.columns.get_level_values(0)     # get labels from first level
dla2level1 = dla2Comparison.columns.get_level_values(1)     # get lables from second leve
dla2Comparison.columns = dla2level0 + '_' + dla2level1      #  Replace old column names w
dla2Comparison.reset_index().head()                         # Reset the index to get them
```

Out[24]:

	Program	DLA2_mean	DLA2_std
0	Intervention	4.500210	0.498272
1	UsualCare	3.863006	0.516134

In [25]:
```
# viewing the difference in Program DLA2 means

dla2Comparison['DLA2_mean'][0] - dla2Comparison['DLA2_mean'][1]
```

Out[25]:
```
0.6372038378288374
```

DLA2 scores represent the 2nd daily living assessment which serves the purpose of expressing any change in scores. If scores from the 2nd DLA assessments are higher than the first, we can assume that the program has had a positive impact on the client. However, if scores decrease, we can assume that the positive impact was either minimal or absent.

Our data shows the Intervention program outperforming the UsualCare program by .63. We will explore the change in DLA scores for both programs to confirm this difference.

In [26]:
```
substanceAbuse2['DLAPctDiff'] = ((substanceAbuse['DLA2']/substanceAbuse['DLA1'])-1)*100

substanceAbuse2.head()
```

Out[26]:

	Admission Date	Program	Age	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	DLA1	DLA2
0	2022-01-13	Intervention	34	F	Other	Depression	Alcohol	2	1	3.69	4.13
1	2022-02-18	Intervention	26	M	NonHispWhite	Trauma	Opioid	0	0	4.22	4.68
2	2022-01-28	Intervention	62	M	NativeAm	Depression	Opioid	0	1	4.17	4.78
3	2022-01-30	Intervention	34	F	NonHispWhite	Depression	Alcohol	0	0	4.11	4.46
4	2022-03-28	UsualCare	46	M	NonHispBlack	Trauma	Opioid	0	1	4.19	4.25

In [27]:
```
# Calculating the average pct change in DLA scores

substanceAbuse2.groupby('Program').agg({'DLAPctDiff':'mean'}).reset_index()
```

Out[27]:

	Program	DLAPctDiff
0	Intervention	15.407081
1	UsualCare	-0.091111

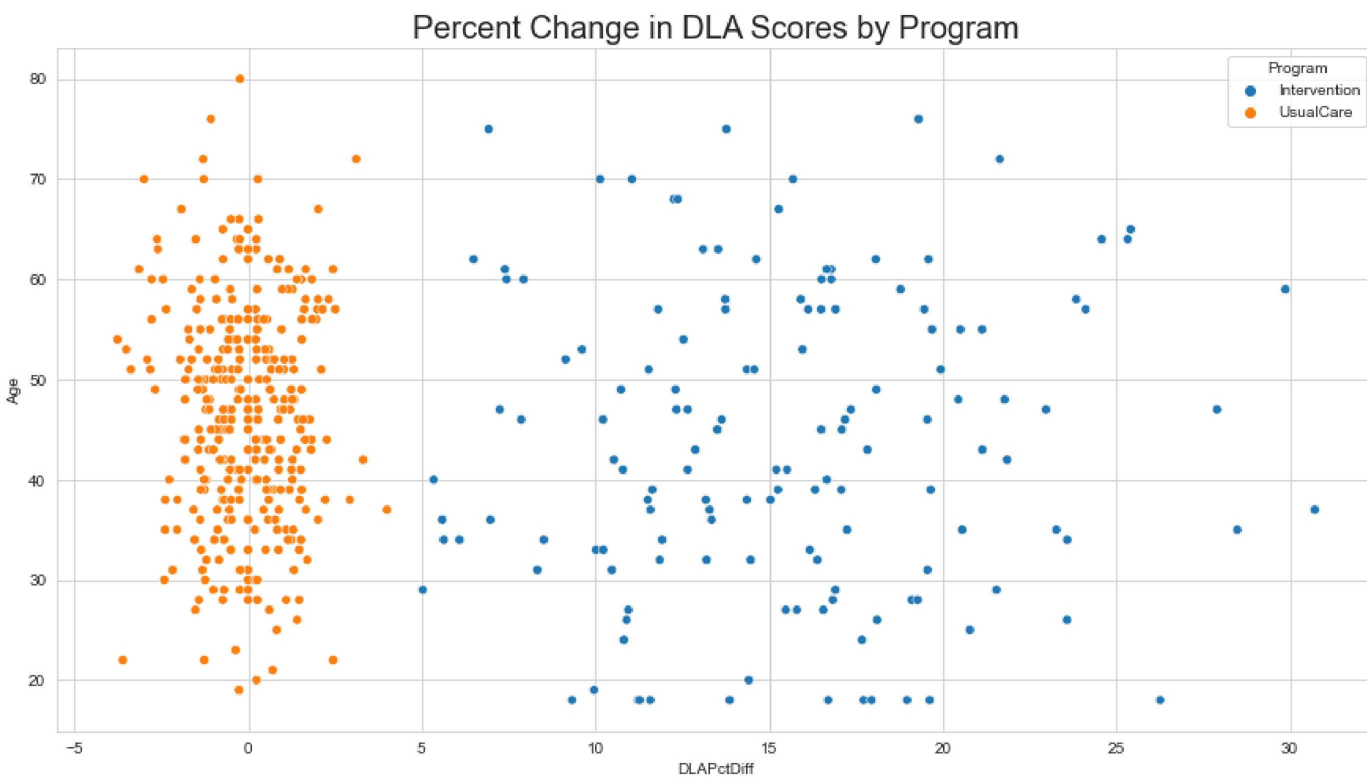
In [28]:
```
# The total pct of people who's scores had a positive change
```

```
totalPositiveDLAChangeByProgram = round(len(substanceAbuse[(substanceAbuse2['DLAPctDiff'

print(f'The percent of observations that had a positive change in DLA scores is {totalPo
```

```
The percent of observations that had a positive change in DLA scores is 0.601
```

In [29]:
```
plt.figure(figsize=(15,8))

sns.scatterplot(substanceAbuse2, x=substanceAbuse2['DLAPctDiff'], y=substanceAbuse2['Age
plt.title(f'Percent Change in DLA Scores by Program', fontsize = 20)
```

Out[29]:
```
Text(0.5, 1.0, 'Percent Change in DLA Scores by Program')
```



The average percent change in DLA scores are +15.4% for Intervention while being -.09% for Usual Care. We can take this information and look at the scatterplot for percent change in DLA scores by program and conclude that Intervention was massively more successful than Usual Care treatment.

2. What are key drivers of different types of primary mental health diagnosis?

In [30]:
```
mentalHealth = substanceAbuse2['MHDx'].value_counts().reset_index().rename(columns = {'i

mentalHealth['Normalize'] = (mentalHealth['Count']/mentalHealth['Count'].sum())*100

mentalHealth
```

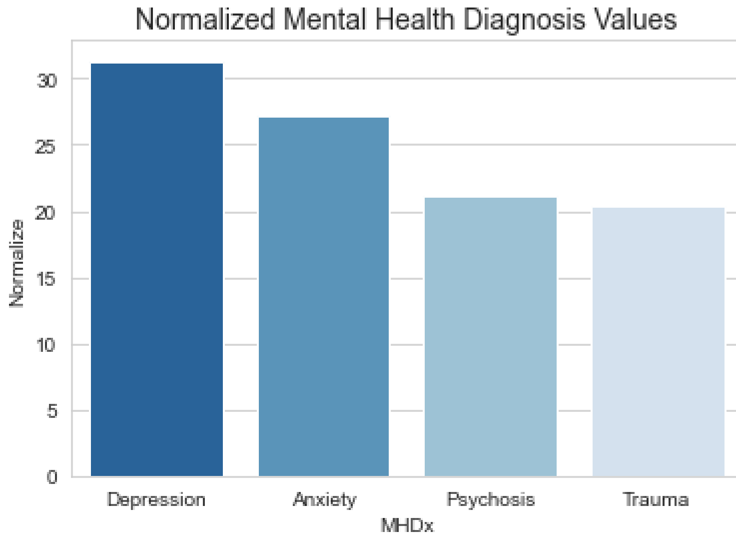
Out[30]:

	MHDx	Count	Normalize
0	Depression	150	31.315240
1	Anxiety	130	27.139875
2	Psychosis	101	21.085595
3	Trauma	98	20.459290

In [31]:
```
sns.barplot(x= mentalHealth['MHDx'], y=mentalHealth['Normalize'], palette= 'Blues_r')
plt.title(f'Normalized Mental Health Diagnosis Values', fontsize= 14)
```

```
Text(0.5, 1.0, 'Normalized Mental Health Diagnosis Values')
```

Out[31]:



The primary mental health diagnosis include depression, anxiety, psychosis and trauma. 58% of clients have been diagnosed with depression or anxiety.

The next thing we'll inspect is mental health diagnosis by program for more insight.

In [32]:

```
usualCareDF = substanceAbuse2[substanceAbuse2['Program']== 'UsualCare']

usualCareDF.head()
```

Out[32]:

	Admission Date	Program	Age	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	DLA1	DLA2
4	2022-03-28	UsualCare	46	M	NonHispBlack	Trauma	Opioid	0	1	4.19	4.25
8	2022-03-26	UsualCare	41	F	Other	Depression	Opioid	1	1	3.95	4.01
9	2022-02-02	UsualCare	33	M	NativeAm	Trauma	None	2	0	4.18	4.16
11	2022-03-17	UsualCare	76	M	NonHispWhite	Psychosis	Opioid	0	1	3.74	3.70
14	2022-03-23	UsualCare	47	M	NonHispWhite	Depression	Stimulant	1	1	4.11	4.06

In [33]:

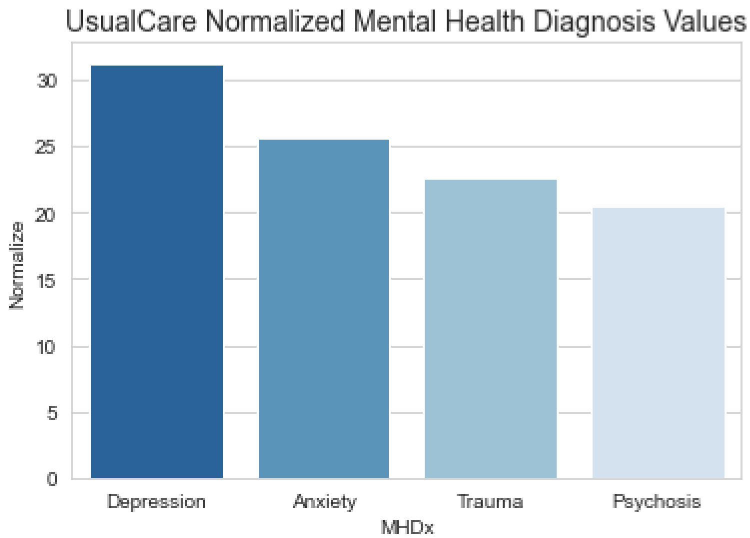
```
usualCareMH = usualCareDF['MHDx'].value_counts().reset_index().rename(columns = {'index'

usualCareMH['Normalize'] = (usualCareMH['Count']/usualCareMH['Count'].sum())*100

usualCareMH
```

Out[33]:

	MHDx	Count	Normalize
0	Depression	105	31.250000
1	Anxiety	86	25.595238
2	Trauma	76	22.619048
3	Psychosis	69	20.535714

In [34]:
```
sns.barplot(x= usualCareMH['MHDx'], y=usualCareMH['Normalize'], palette= 'Blues_r')
plt.title(f'UsualCare Normalized Mental Health Diagnosis Values', fontsize= 14)
```

Out[34]: Text(0.5, 1.0, 'UsualCare Normalized Mental Health Diagnosis Values')



In [35]:
```
interventionCareDF = substanceAbuse2[substanceAbuse['Program'] == 'Intervention']

interventionCareDF.head()
```

Out[35]:

	Admission Date	Program	Age	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	DLA1	DLA2
0	2022-01-13	Intervention	34	F	Other	Depression	Alcohol	2	1	3.69	4.13
1	2022-02-18	Intervention	26	M	NonHispWhite	Trauma	Opioid	0	0	4.22	4.68
2	2022-01-28	Intervention	62	M	NativeAm	Depression	Opioid	0	1	4.17	4.78
3	2022-01-30	Intervention	34	F	NonHispWhite	Depression	Alcohol	0	0	4.11	4.46
5	2022-02-17	Intervention	51	M	NonHispWhite	Anxiety	Opioid	1	0	3.55	4.06

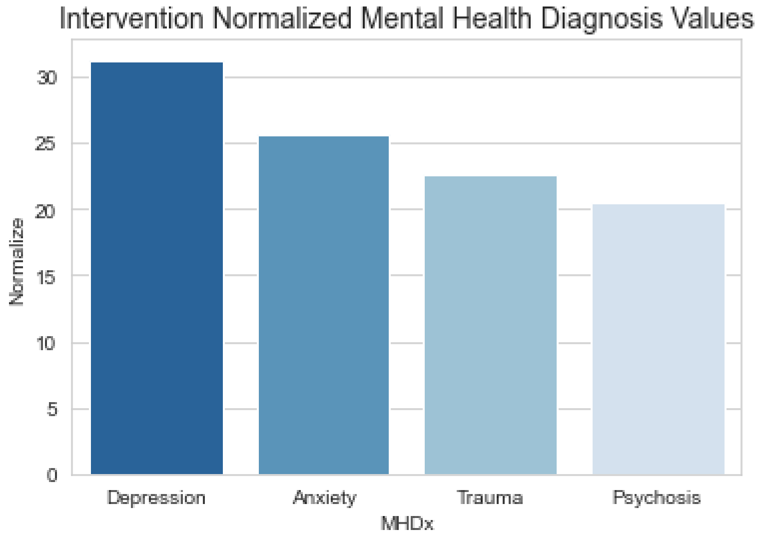
In [36]:
```
interventionMH = interventionCareDF['MHDx'].value_counts().reset_index().rename(columns=

interventionMH['Normalized'] = (interventionMH['Count']/interventionMH['Count'].sum())*1

interventionMH
```

Out[36]:

	MHDx	Count	Normalized
0	Depression	45	31.468531
1	Anxiety	44	30.769231
2	Psychosis	32	22.377622
3	Trauma	22	15.384615

In [37]:
```
sns.barplot(x= usualCareMH['MHDx'], y=usualCareMH['Normalize'], palette= 'Blues_r')
plt.title(f'Intervention Normalized Mental Health Diagnosis Values', fontsize= 14)
```

Out[37]: Text(0.5, 1.0, 'Intervention Normalized Mental Health Diagnosis Values')



Now that we've inspected the mental health diagnosis for both programs, we can say that depression and anxiety are the two key primary drivers for mental health diagnosis in both programs.

3.Demographic analysis about different types of primary mental health diagnosis?

In [38]:
```
substanceAbuse2.columns
```

Out[38]:
```
Index(['Admission Date', 'Program', 'Age', 'Gender', 'RaceEthnicity', 'MHDx',
       'SUDx', 'MedDx', 'PsychAdmit', 'DLA1', 'DLA2', 'Age Group',
       'DLAPctDiff'],
      dtype='object')
```

In [39]:
```
# creating demorgraphics variable for easy slicing
# creating a dataframe with demographic info and other interesting variables to explore

demographics = [ 'Gender', 'RaceEthnicity', 'MHDx',
       'SUDx', 'MedDx', 'PsychAdmit', 'Age Group']

demographicsDf = substanceAbuse2[demographics]

demographicsDf
```

Out[39]:

	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	Age Group
0	F	Other	Depression	Alcohol	2	1	30's
1	M	NonHispWhite	Trauma	Opioid	0	0	Under 30
2	M	NativeAm	Depression	Opioid	0	1	60's
3	F	NonHispWhite	Depression	Alcohol	0	0	30's
4	M	NonHispBlack	Trauma	Opioid	0	1	40's
...	...	...	...	...	...	...	...
474	M	Other	Psychosis	Stimulant	1	1	50's
475	F	NativeAm	Depression	Alcohol	1	0	40's
476	M	NativeAm	Psychosis	None	2	1	30's
477	M	Other	Anxiety	Opioid	0	0	60's
478	F	Other	Psychosis	None	2	1	60's

479 rows × 7 columns

In [40]:
```
demographicsDf.groupby('RaceEthnicity')['MHDx'].value_counts().unstack()
```

Out[40]:

MHDx	Anxiety	Depression	Psychosis	Trauma
RaceEthnicity				
Hispanic	3	8	2	6
NativeAm	30	47	19	23
NonHispBlack	6	7	4	5
NonHispWhite	32	48	32	22
Other	59	40	44	42

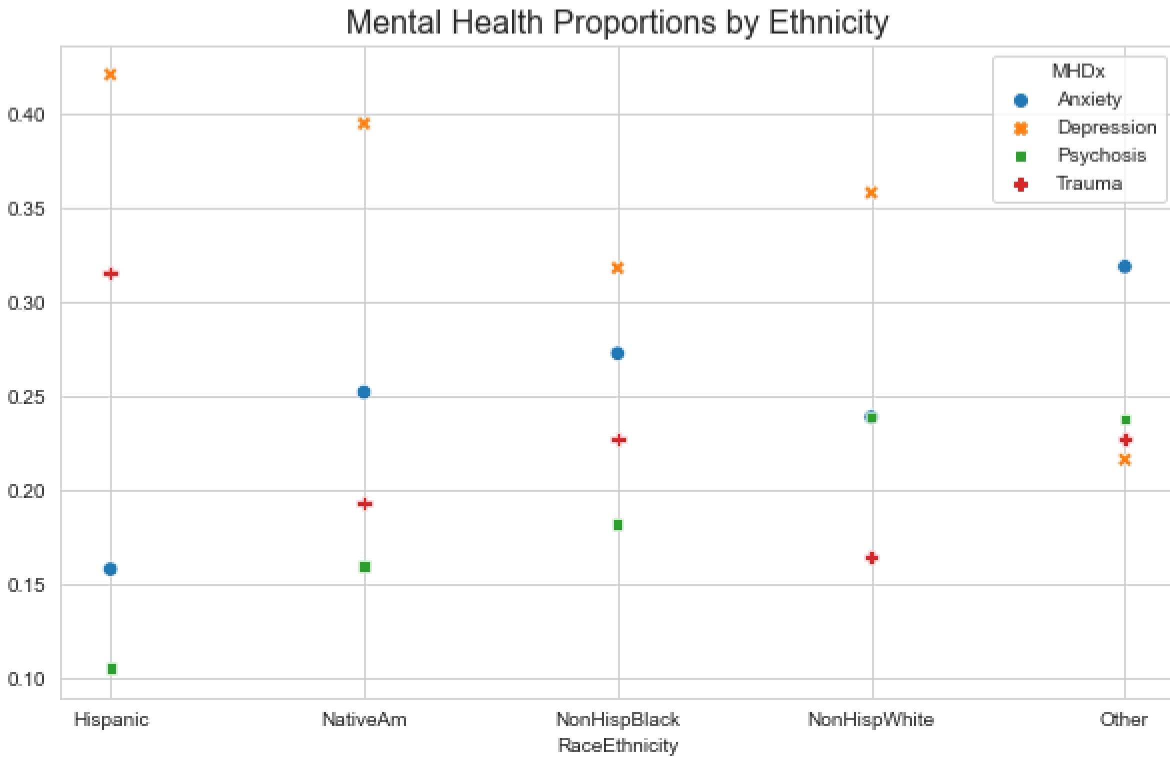
In [41]:
```
raceDf = demographicsDf.groupby('RaceEthnicity')['MHDx'].value_counts(normalize=True).un

# duplicate dataframe for scatterplot
raceDf2 = demographicsDf.groupby('RaceEthnicity')['MHDx'].value_counts(normalize=True).u



raceDf
```

Out[41]:

MHDx	Anxiety	Depression	Psychosis	Trauma
RaceEthnicity				
Hispanic	0.157895	0.421053	0.105263	0.315789
NativeAm	0.252101	0.394958	0.159664	0.193277
NonHispBlack	0.272727	0.318182	0.181818	0.227273
NonHispWhite	0.238806	0.358209	0.238806	0.164179
Other	0.318919	0.216216	0.237838	0.227027

In [42]:
```
plt.figure(figsize=(10,6))

sns.scatterplot(raceDf2, s=55)
plt.title(f'Mental Health Proportions by Ethnicity', fontsize= 16)
```

Out[42]:
```
Text(0.5, 1.0, 'Mental Health Proportions by Ethnicity')
```



As we look at the graph, we can see that all ethnicities were affected the most by depression besides the group labeled Other, who's leading mental health diagnosis was anxiety. The mental health diagnosis trend transitions from depression to anxiety as the top factors. However, this is not the case for the hispanic poppulation who's 2nd leading diagnosis is trauma.

In [43]:
```
demographicsDf.groupby('Gender')['MHDx'].value_counts().unstack()
```

Out[43]:

MHDx	Anxiety	Depression	Psychosis	Trauma
Gender				
F	47	55	38	39
M	83	95	63	59

In [44]:
```
genderDf = demographicsDf.groupby('Gender')['MHDx'].value_counts(normalize=True).unstack

genderDf
```

Out[44]:

MHDx	Anxiety	Depression	Psychosis	Trauma
Gender				
F	0.262570	0.307263	0.212291	0.217877
M	0.276667	0.316667	0.210000	0.196667

In [45]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(genderDf, s=55)
plt.title(f'Mental Health Proportions by Gender', fontsize= 16)
```

Out[45]:
```
Text(0.5, 1.0, 'Mental Health Proportions by Gender')
```

The mental health diagnosis for genders reflects each other and displays little difference in proportions.

Inspecting Abnormal Mental Health Conditions

There are a number of peculiar instances in the dataset that represent what we should label as 'at-risk' individuals. These observations include clients that were admitted to pysch more than one time.

In [46]:

```
demographicsDf
```

Out[46]:

	Gender	RaceEthnicity	MHDx	SUDx	MedDx	PsychAdmit	Age Group
0	F	Other	Depression	Alcohol	2	1	30's
1	M	NonHispWhite	Trauma	Opioid	0	0	Under 30
2	M	NativeAm	Depression	Opioid	0	1	60's
3	F	NonHispWhite	Depression	Alcohol	0	0	30's
4	M	NonHispBlack	Trauma	Opioid	0	1	40's
...	...	...	...	...	...	...	...
474	M	Other	Psychosis	Stimulant	1	1	50's
475	F	NativeAm	Depression	Alcohol	1	0	40's
476	M	NativeAm	Psychosis	None	2	1	30's
477	M	Other	Anxiety	Opioid	0	0	60's
478	F	Other	Psychosis	None	2	1	60's

479 rows × 7 columns

In [47]:

```
# creating an abnormal dataframe

mhxAndSudx = ['MHDx', 'SUDx']
demographics2 = [ 'Gender', 'RaceEthnicity',
```

```
        'MedDx', 'PsychAdmit', 'Age']
abnormalDf = substanceAbuse2[mhxAndSudx + demographics2]

abnormalDf
```

Out[47]:

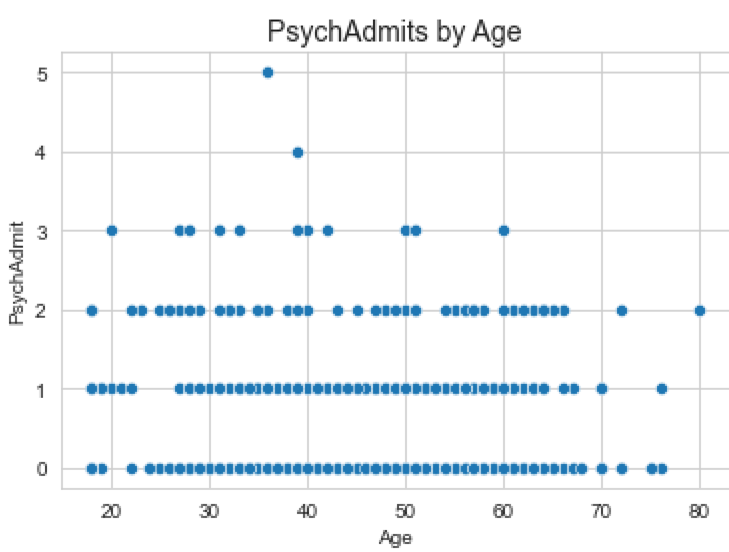
	MHDx	SUDx	Gender	RaceEthnicity	MedDx	PsychAdmit	Age
0	Depression	Alcohol	F	Other	2	1	34
1	Trauma	Opioid	M	NonHispWhite	0	0	26
2	Depression	Opioid	M	NativeAm	0	1	62
3	Depression	Alcohol	F	NonHispWhite	0	0	34
4	Trauma	Opioid	M	NonHispBlack	0	1	46
...	...	...	...	...	...	...	...
474	Psychosis	Stimulant	M	Other	1	1	58
475	Depression	Alcohol	F	NativeAm	1	0	47
476	Psychosis	None	M	NativeAm	2	1	39
477	Anxiety	Opioid	M	Other	0	0	63
478	Psychosis	None	F	Other	2	1	66

479 rows × 7 columns

In [48]:

```
sns.scatterplot(data = abnormalDf,  x= abnormalDf['Age'],y= abnormalDf['PsychAdmit'])
plt.title(f'PsychAdmits by Age', fontsize = 14)
```

Out[48]: Text(0.5, 1.0, 'PsychAdmits by Age')



In [49]:

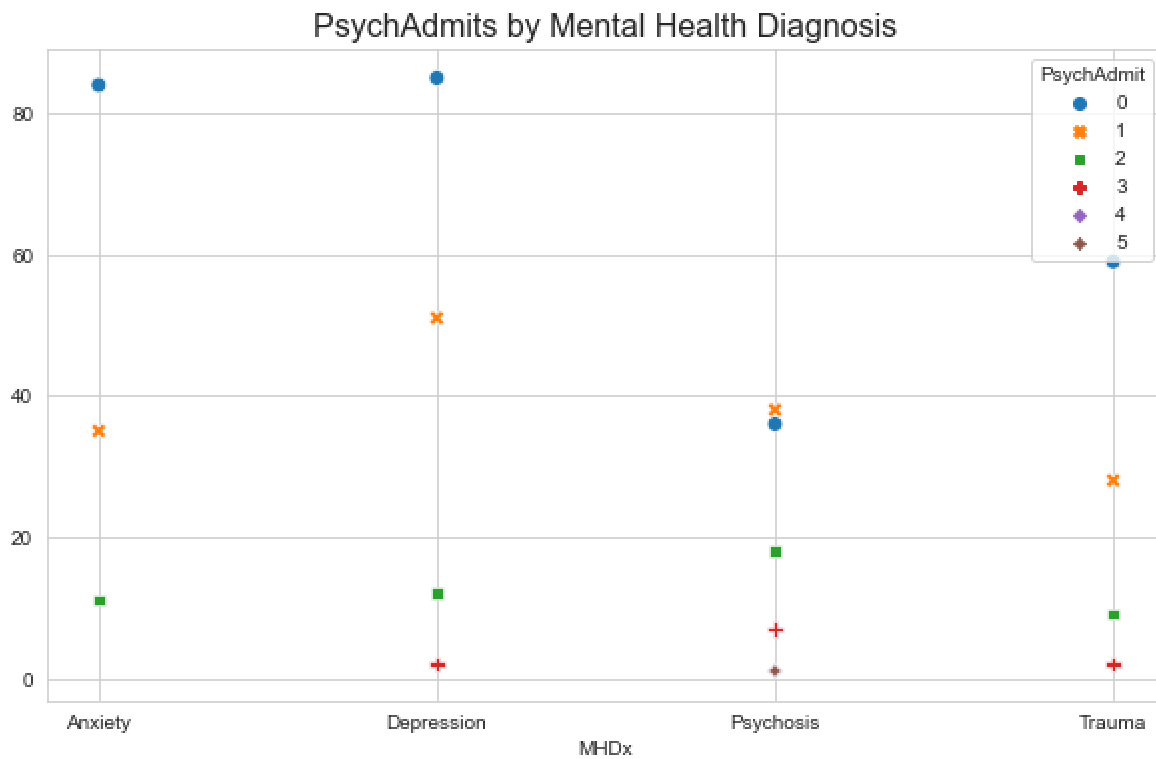
```
mHPsychAdmits = abnormalDf.groupby('MHDx')['PsychAdmit'].value_counts().unstack()
abnormalDf.groupby('MHDx')['PsychAdmit'].value_counts().unstack().style.highlight_max()
```

Out[49]:

PsychAdmit	0	1	2	3	4	5
MHDx						
Anxiety	84.000000	35.000000	11.000000	nan	nan	nan
Depression	85.000000	51.000000	12.000000	2.000000	nan	nan
Psychosis	36.000000	38.000000	18.000000	7.000000	1.000000	1.000000
Trauma	59.000000	28.000000	9.000000	2.000000	nan	nan

In [50]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(mHPsychAdmits,s=60)
plt.title(f'PsychAdmits by Mental Health Diagnosis', fontsize=16)
```

Out[50]: `Text(0.5, 1.0, 'PsychAdmits by Mental Health Diagnosis')`



We can observe Pscychosis as the most significant contributor to clients with multiple Psych admits, which makes sense. Considering that depression is the second leading cause only brings more attention to it needing to be addressed.

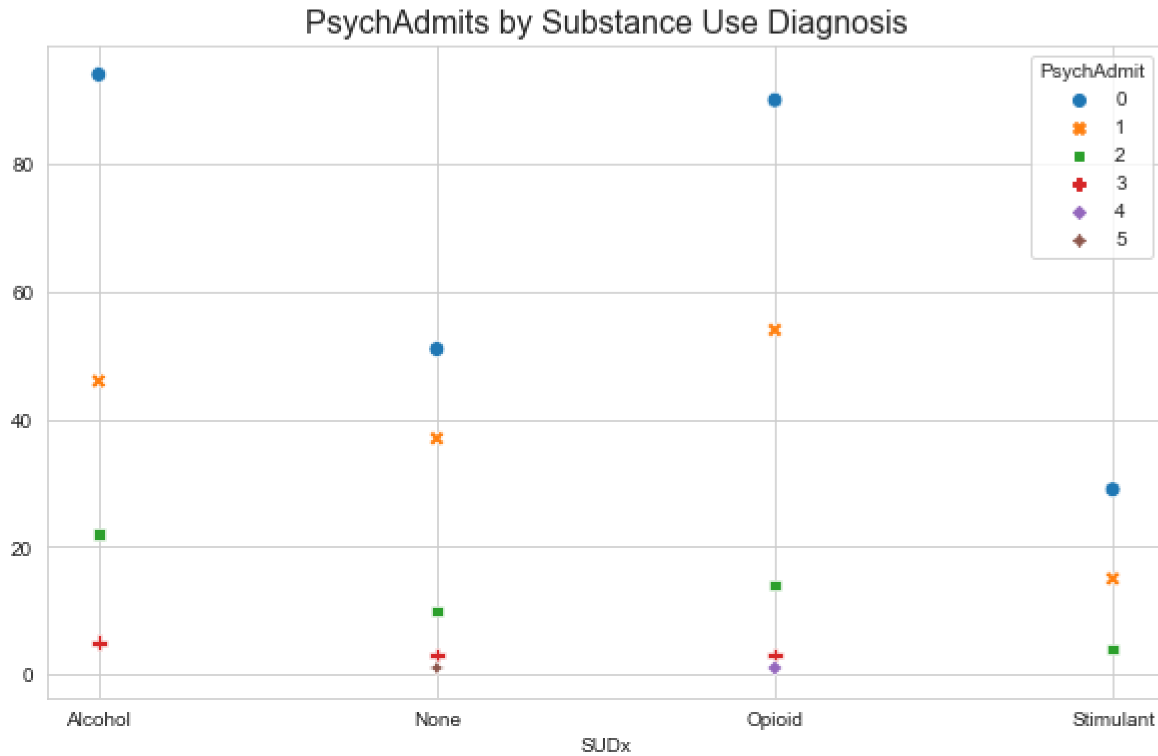
In [51]:
```
sudxPsychAdmits = abnormalDf.groupby('SUDx')['PsychAdmit'].value_counts().unstack()
abnormalDf.groupby('SUDx')['PsychAdmit'].value_counts().unstack().style.highlight_max()
```

Out[51]:

PsychAdmit	0	1	2	3	4	5
SUDx						
Alcohol	94.000000	46.000000	22.000000	5.000000	nan	nan
None	51.000000	37.000000	10.000000	3.000000	nan	1.000000
Opioid	90.000000	54.000000	14.000000	3.000000	1.000000	nan
Stimulant	29.000000	15.000000	4.000000	nan	nan	nan

In [52]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(sudxPsychAdmits,s=60)
plt.title(f'PsychAdmits by Substance Use Diagnosis', fontsize=16)
```

Out[52]: `Text(0.5, 1.0, 'PsychAdmits by Substance Use Diagnosis')`



Alcohol is the leading substance that is associated with recurring psych visits.

Recommendations

1. Create a practical communications and relationship building class.

1. Enroll all clients into a health and wellness program which provides exercise twice a day for 45mins and stress reduction techniques such as meditation, yoga and community walks. This program should be mandatory.

1. Create bi-weekly community cookout events that include sports, board games and local music artists to help give clients an opportunity to restore their relationships with their family and develop new connections.

1. Increase the employee to client ratio by hiring more employees or reducing the max number of beds for clients for the Usual Care program.