In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from feature_selector.feature_selector import FeatureSelector
```

executed in 48.5s, finished 21:00:26 2019-01-29

In [2]:

```
train = pd.read_csv('train (1).csv')
```

executed in 188ms, finished 21:00:26 2019-01-29

In [3]:

```
train.drop(['survey_date','surveyid'],inplace=True,axis=1)
```

executed in 98ms, finished 21:00:26 2019-01-29

In [4]:

```
train.info()
```

executed in 417ms, finished 21:00:26 2019-01-29

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 73 columns):
village                    1143 non-null int64
femaleres                  1143 non-null int64
age                        1143 non-null float64
married                    1143 non-null int64
children                   1143 non-null int64
hhsize                     1143 non-null int64
edu                        1143 non-null int64
hh_children                1143 non-null int64
hh_totalmembers            809 non-null float64
cons_nondurable            1143 non-null float64
asset_livestock            1143 non-null float64
asset_durable              1143 non-null float64
asset_phone                1143 non-null float64
asset_savings              1143 non-null float64
asset_land_owned_total     1143 non-null float64
asset_niceroof             1143 non-null int64
cons_allfood               1143 non-null float64
cons_ownfood               1143 non-null float64
cons_alcohol               1100 non-null float64
cons_tobacco               1123 non-null float64
cons_med_total             1143 non-null float64
cons_med_children          724 non-null float64
cons_ed                    1143 non-null float64
cons_social                1143 non-null float64
cons_other                 1143 non-null float64
ent_wagelabor              1143 non-null int64
ent_ownfarm                1143 non-null int64
ent_business               1143 non-null int64
ent_nonagbusiness          1143 non-null int64
ent_employees              1143 non-null int64
ent_nonag_revenue          1143 non-null float64
ent_nonag_flowcost         1143 non-null float64
ent_farmrevenue            1143 non-null float64
ent_farmexpenses           1143 non-null float64
ent_animalstockrev         1143 non-null float64
ent_total_cost             1143 non-null float64
fs_adskipm_often           1143 non-null float64
fs_adwholed_often          1143 non-null float64
fs_chskipm_often           727 non-null float64
fs_chwholed_often          727 non-null float64
fs_meat                    809 non-null float64
fs_enoughtom               809 non-null float64
fs_sleephun                809 non-null float64
med_expenses_hh_ep         450 non-null float64
med_expenses_sp_ep         265 non-null float64
med_expenses_child_ep      543 non-null float64
med_portion_sickinjured    809 non-null float64
med_port_sick_child        727 non-null float64
med_afford_port            720 non-null float64
med_sickdays_hhave         809 non-null float64
med_healthconsult          720 non-null float64
med_vacc_newborns          1143 non-null int64
med_child_check            1143 non-null int64
```

```
med_u5_deaths              59 non-null float64
ed_expenses                680 non-null float64
ed_expenses_perkid         680 non-null float64
ed_schoolattend            680 non-null float64
ed_sch_missedpc            676 non-null float64
ed_work_act_pc             572 non-null float64
labor_primary              1143 non-null int64
wage_expenditures          1143 non-null int64
durable_investment         1143 non-null float64
nondurable_investment      1143 non-null float64
given_mpesa                1143 non-null int64
amount_given_mpesa         1143 non-null float64
received_mpesa             1143 non-null int64
amount_received_mpesa      1143 non-null float64
net_mpesa                  1143 non-null float64
saved_mpesa                1143 non-null int64
amount_saved_mpesa         1143 non-null float64
early_survey               1143 non-null int64
depressed                  1143 non-null int64
day_of_week                1143 non-null int64
dtypes: float64(50), int64(23)
memory usage: 651.9 KB
```

In [5]:

```
nul_col=[[col,train[col].isnull().sum()] for col in train.columns if train[col].isnull().
```

executed in 240ms, finished 21:00:27 2019-01-29

In [6]:

```
print(nul_col)
```

executed in 60ms, finished 21:00:27 2019-01-29

```
[['hh_totalmembers', 334], ['cons_alcohol', 43], ['cons_tobacco', 20], ['con
s_med_children', 419], ['fs_chskipm_often', 416], ['fs_chwholed_often', 41
6], ['fs_meat', 334], ['fs_enoughtom', 334], ['fs_sleephun', 334], ['med_exp
enses_hh_ep', 693], ['med_expenses_sp_ep', 878], ['med_expenses_child_ep', 6
00], ['med_portion_sickinjured', 334], ['med_port_sick_child', 416], ['med_a
fford_port', 423], ['med_sickdays_hhave', 334], ['med_healthconsult', 423],
['med_u5_deaths', 1084], ['ed_expenses', 463], ['ed_expenses_perkid', 463],
['ed_schoolattend', 463], ['ed_sch_missedpc', 467], ['ed_work_act_pc', 571]]
```

In [7]:

```
un=[col for col in train.columns if  train[col].isnull().sum()/1143 > 0.2]
```

executed in 330ms, finished 21:00:27 2019-01-29

In [8]:

```
clean_col=list(set(train.columns)-set(un))
```

executed in 162ms, finished 21:00:27 2019-01-29

In [9]:

```
len(clean_col)
```

executed in 239ms, finished 21:00:27 2019-01-29

Out[9]:

```
52
```

In [10]:

```
tr=train[clean_col]
```

executed in 177ms, finished 21:00:28 2019-01-29

In [11]:

```
tr.nunique()
```

executed in 298ms, finished 21:00:28 2019-01-29

Out[11]:

```
fs_adwholed_often           5
femaleres                   2
cons_nondurable           808
saved_mpesa                 2
ent_farmrevenue           309
given_mpesa                 2
edu                        18
cons_allfood              763
asset_phone                77
received_mpesa              2
fs_adskipm_often            5
cons_other                549
ent_nonag_revenue         110
ent_wagelabor               2
asset_land_owned_total     61
hhsize                     12
asset_livestock           274
asset_durable             586
ent_employees               5
amount_received_mpesa      28
nondurable_investment     767
amount_saved_mpesa         41
early_survey                2
ent_nonag_flowcost        156
labor_primary               2
asset_niceroof              2
ent_total_cost            704
day_of_week                 7
ent_business                2
village                   241
cons_ownfood              466
cons_social               334
ent_animalstockrev        224
ent_nonagbusiness           2
med_vacc_newborns           1
hh_children                11
amount_given_mpesa         12
ent_farmexpenses          575
durable_investment        794
age                        99
net_mpesa                  38
cons_med_total            102
cons_tobacco               36
asset_savings              80
depressed                   2
children                   11
cons_alcohol               35
ent_ownfarm                 2
cons_ed                   263
wage_expenditures           3
med_child_check             1
married                     2
dtype: int64
```

In [12]:

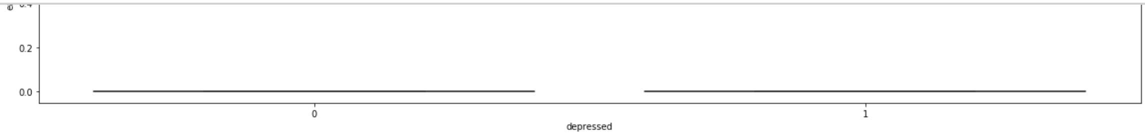
```
cat_col = [col for col in tr.columns if tr[col].nunique()<12]
num_col = [col for col in tr.columns if tr[col].nunique()>12]
```

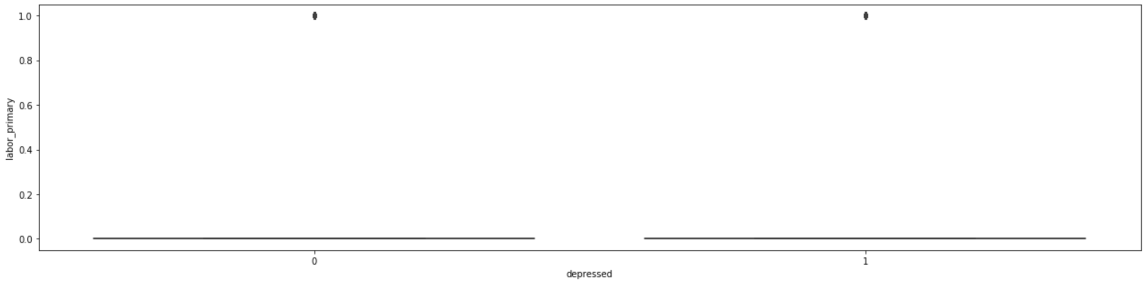
executed in 271ms, finished 21:00:28 2019-01-29

In [13]:

```
def plot_bar(data, cols, col_x = None):
    for col in cols:
        plt.figure(figsize=(22,5))
        sns.boxplot(col_x, y=col, data=data)
        plt.xlabel(col_x) # Set text for the x axis
        plt.ylabel(col)# Set text for y axis
        plt.show()

plot_bar(data=tr,cols=cat_col,col_x='depressed')
```

executed in 20.9s, finished 21:00:49 2019-01-29



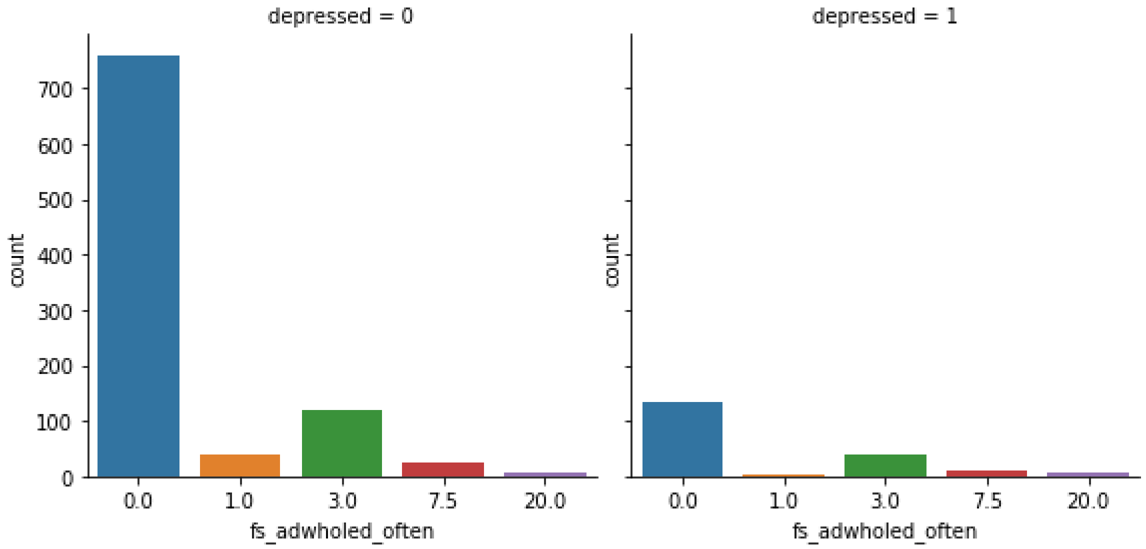


In [14]:

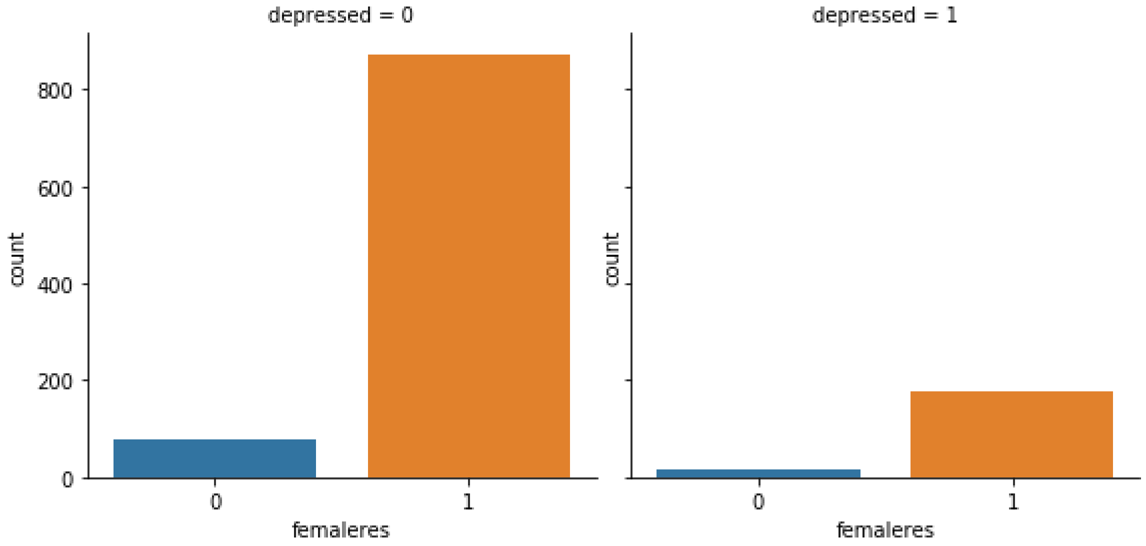
```
def plot_bar(data, cols,hue='depressed'):
    for col in cols:
        plt.figure(figsize=(22,5))
        g = sns.factorplot(x=col, col=hue,
data=data, kind="count");
        plt.xlabel(col) # Set text for the x axis
        plt.ylabel('count')# Set text for y axis
        plt.show()

plot_bar(data=tr,cols=cat_col)
```

executed in 32.3s, finished 21:01:22 2019-01-29

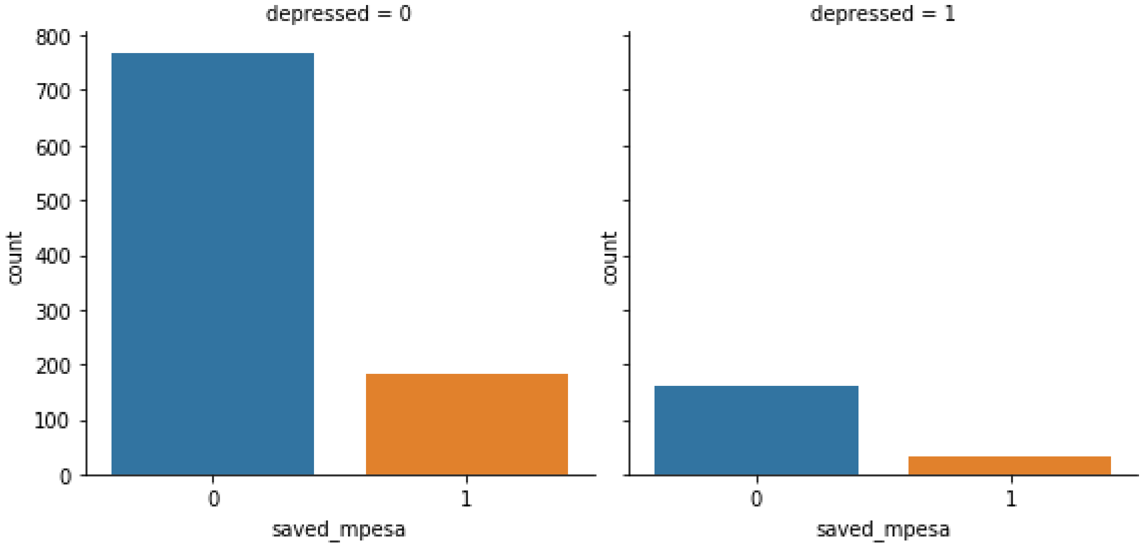
```
<matplotlib.figure.Figure at 0x1ff654908d0>
```



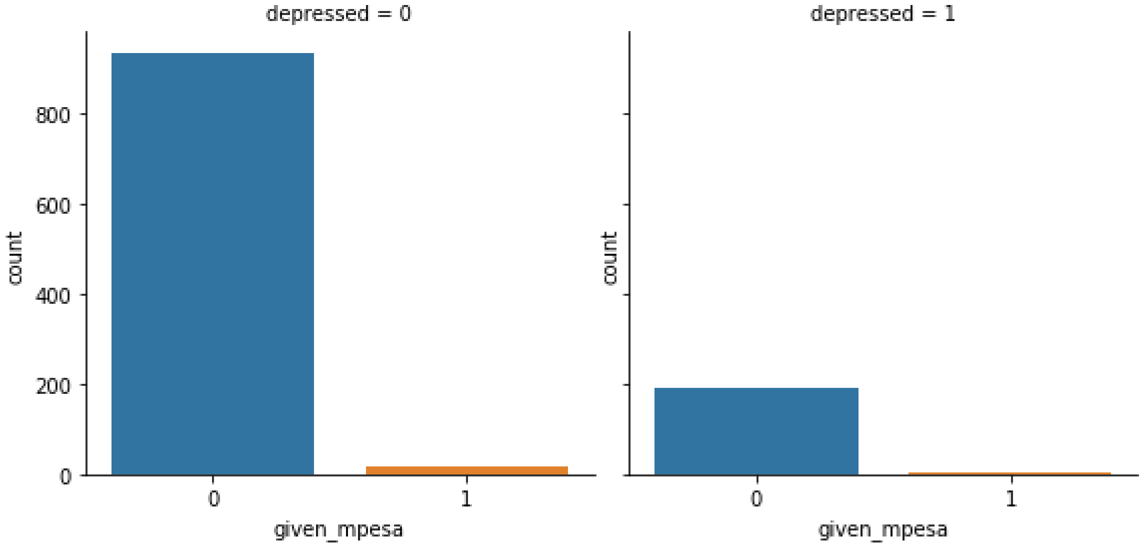
```
<matplotlib.figure.Figure at 0x1ff656b2630>
```



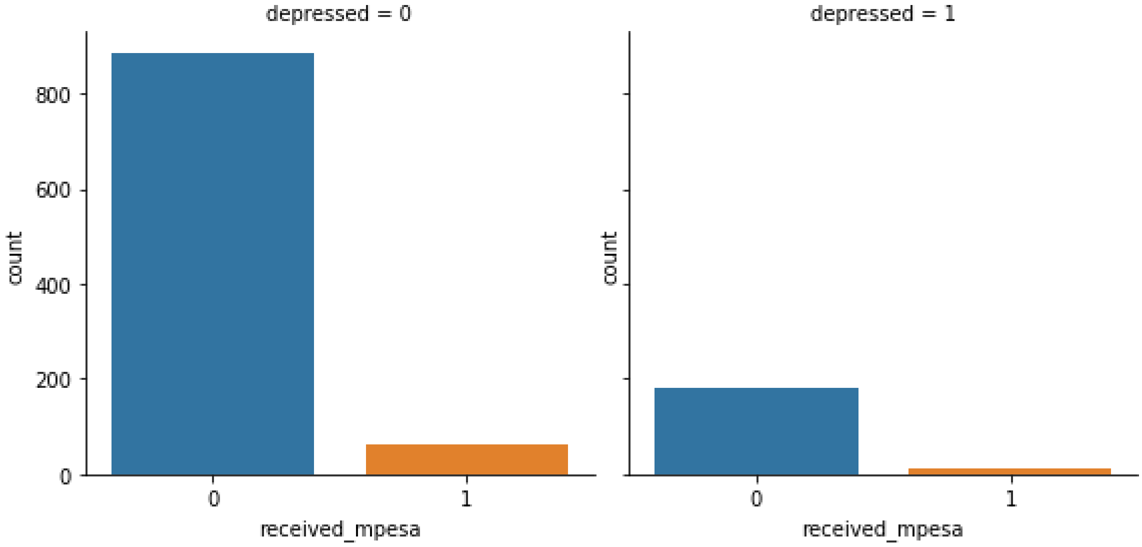
```
<matplotlib.figure.Figure at 0x1ff63fefa90>
```


depressed = 0
depressed = 1
count
800
700
600
500
400
300
200
100
0
0
1
saved_mpesa
count
0
1
saved_mpesa

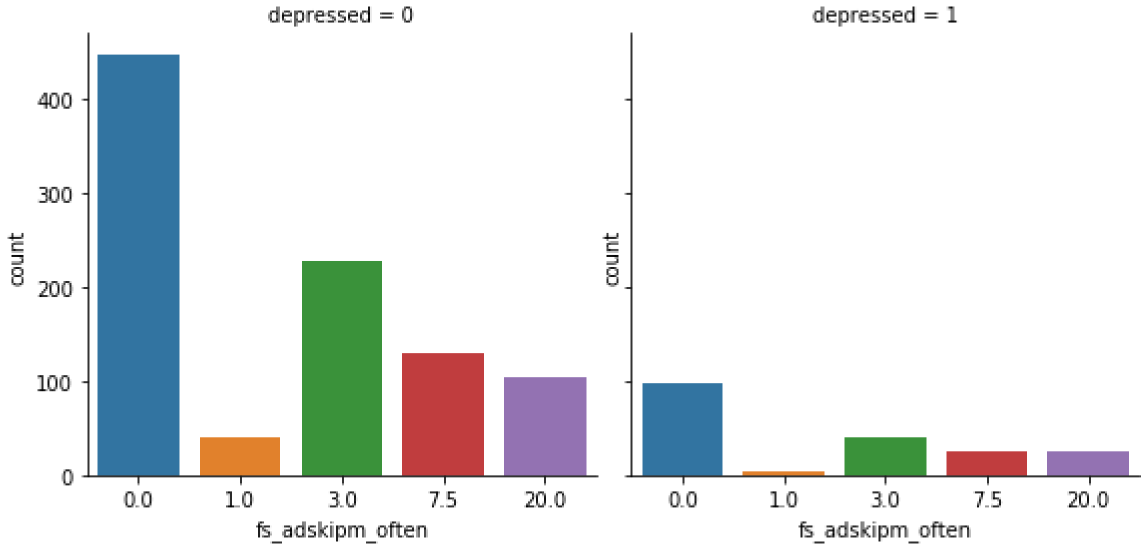
```
<matplotlib.figure.Figure at 0x1ff64c9f358>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
1
given_mpesa
count
0
1
given_mpesa

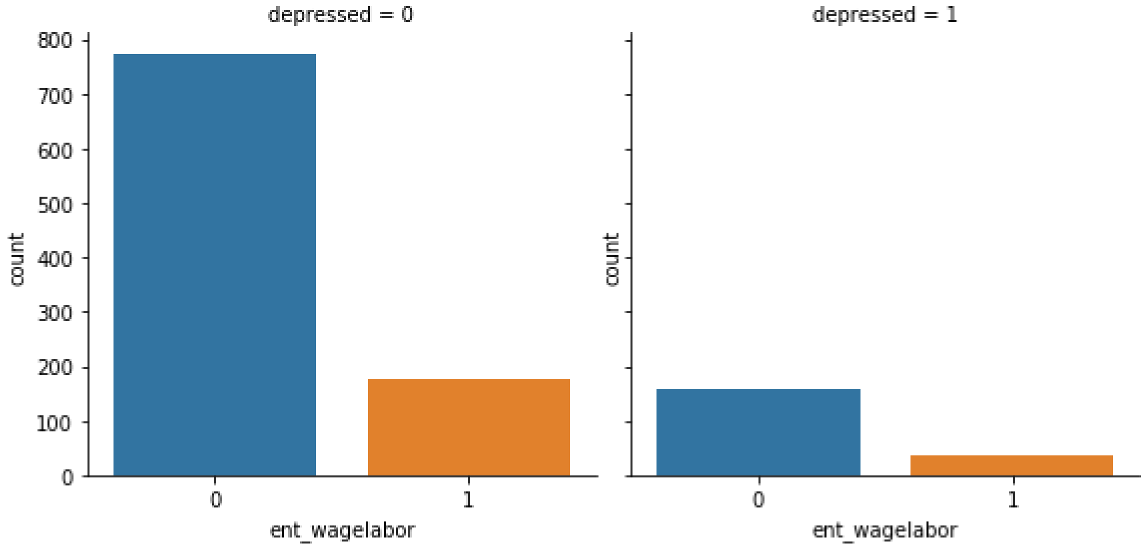
```
<matplotlib.figure.Figure at 0x1ff63fc3a58>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
1
received_mpesa
count
0
1
received_mpesa

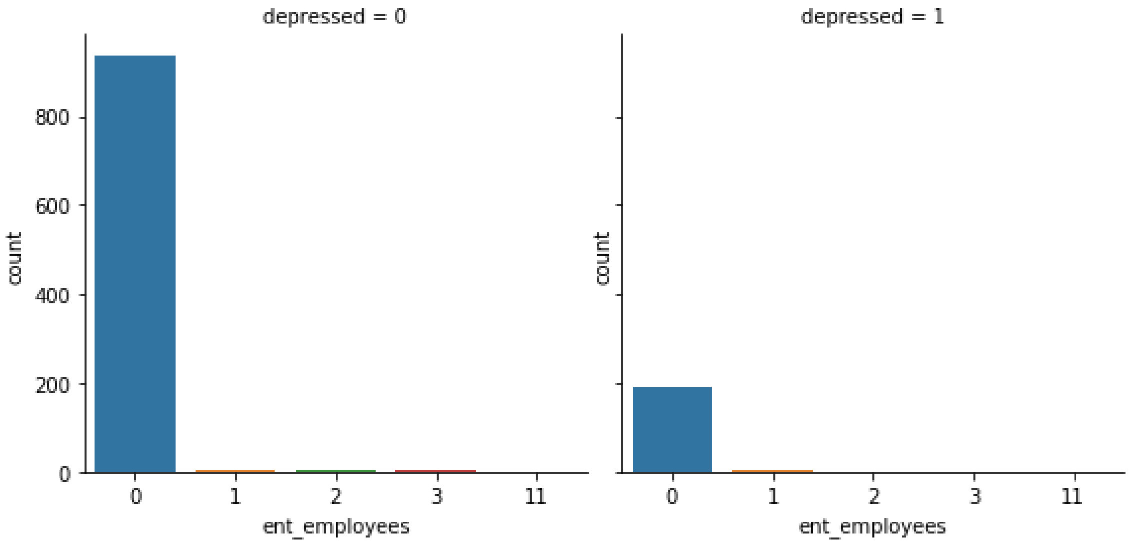
```
<matplotlib.figure.Figure at 0x1ff640ec710>
```


depressed = 0
depressed = 1
count
0
100
200
300
400
0.0
1.0
3.0
7.5
20.0
fs_adskipm_often

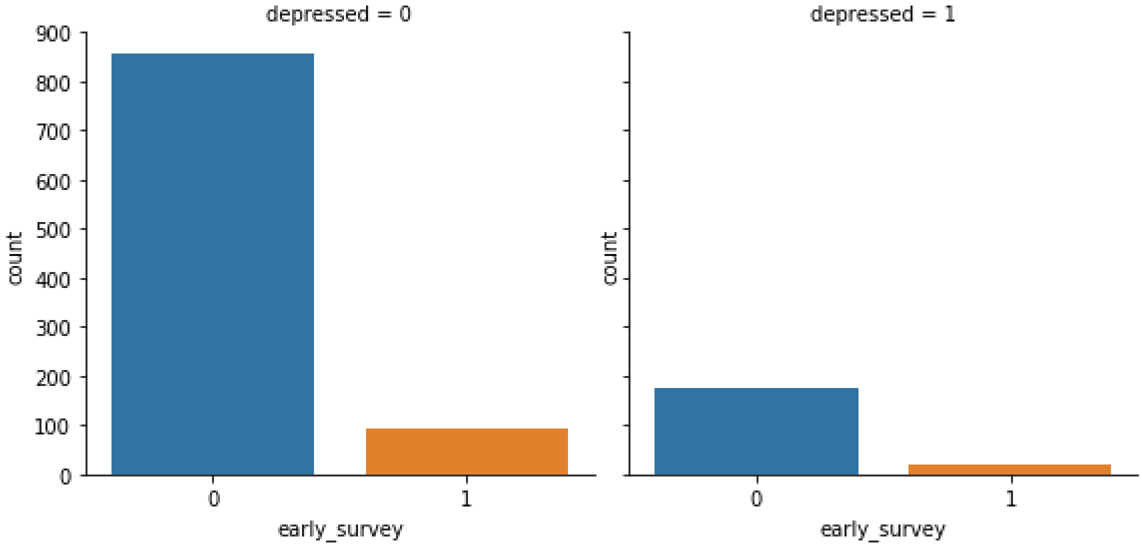
```
<matplotlib.figure.Figure at 0x1ff6412f048>
```


depressed = 0
depressed = 1
count
0
100
200
300
400
500
600
700
800
0
1
ent_wagelabor

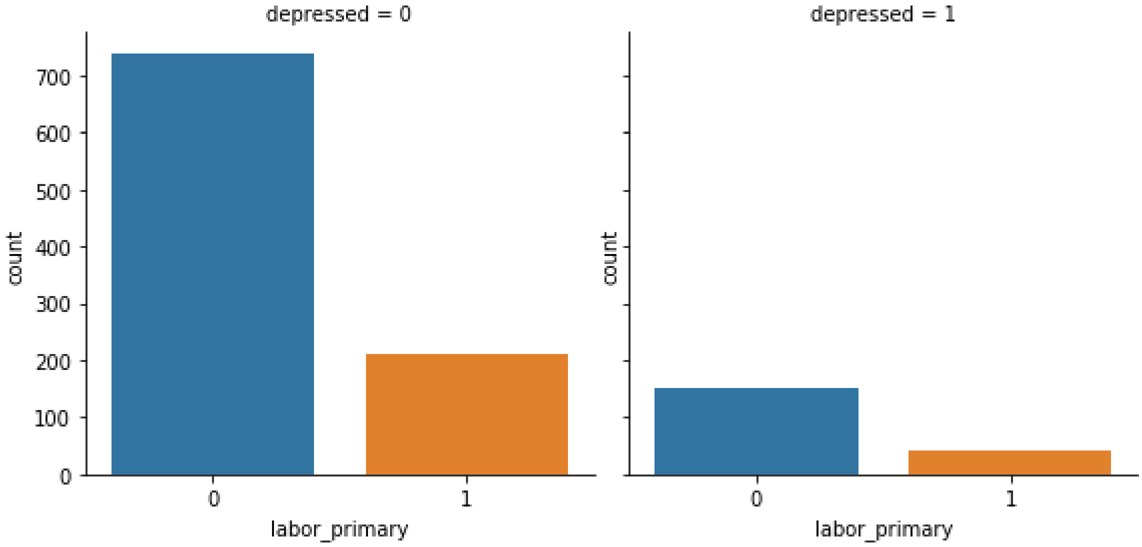
```
<matplotlib.figure.Figure at 0x1ff64102b70>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
1
2
3
11
ent_employees

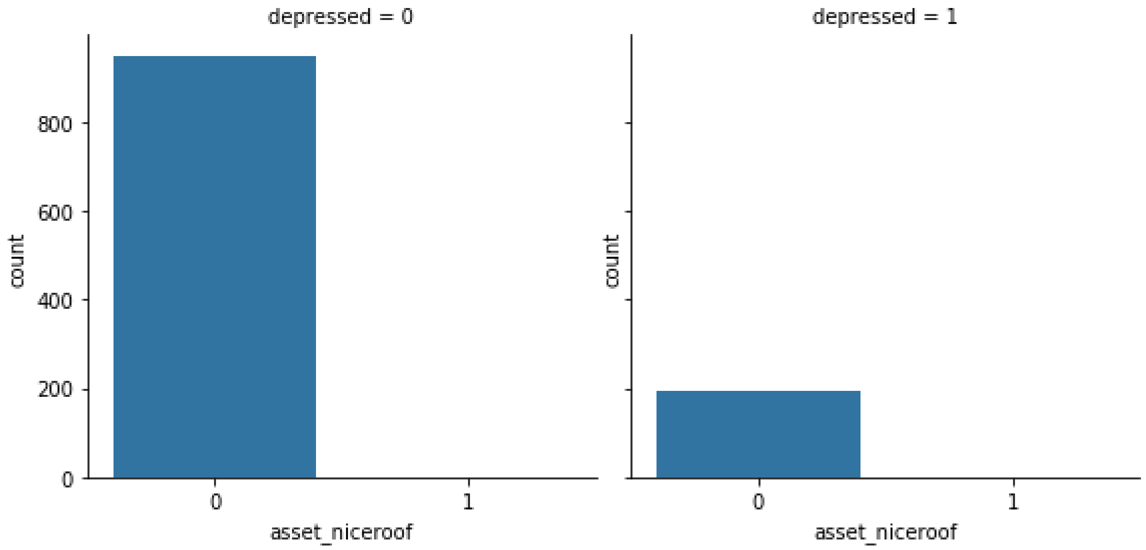
```
<matplotlib.figure.Figure at 0x1ff64005240>
```


depressed = 0
depressed = 1
count
900
800
700
600
500
400
300
200
100
0
0
1
early_survey

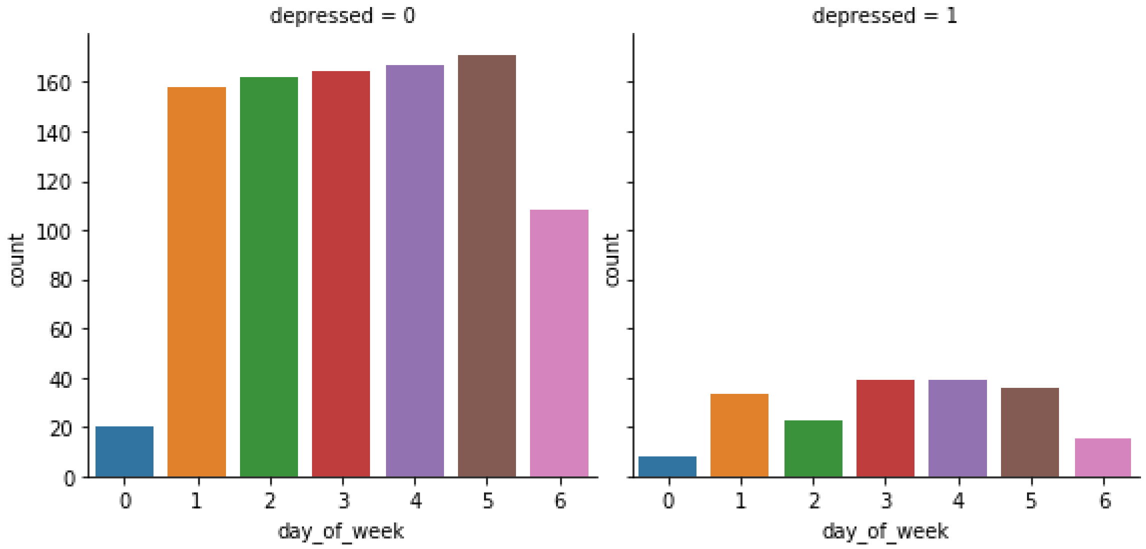
```
<matplotlib.figure.Figure at 0x1ff63e61780>
```


depressed = 0
depressed = 1
count
700
600
500
400
300
200
100
0
0
1
labor_primary

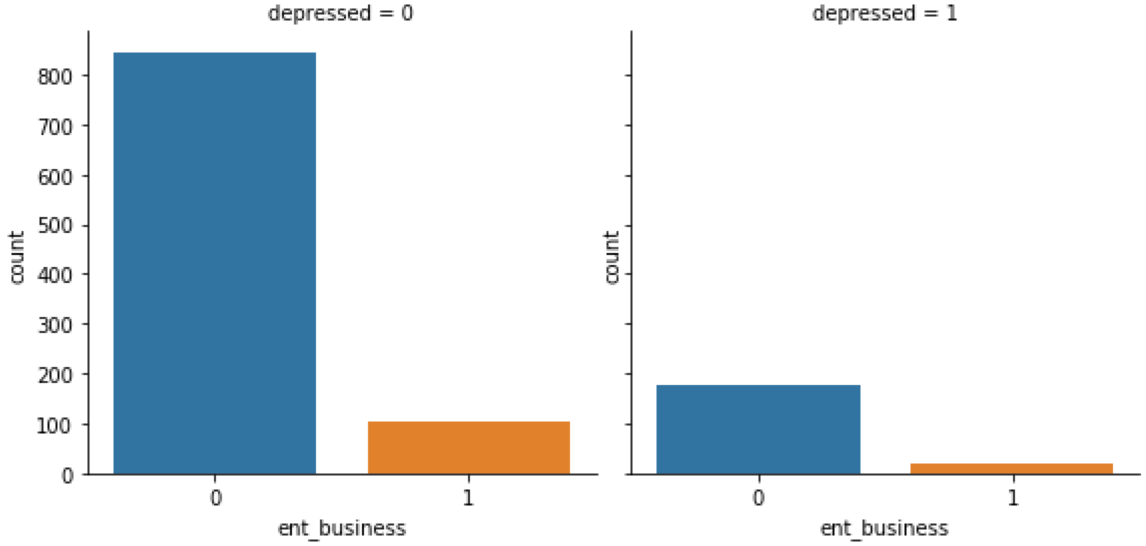
```
<matplotlib.figure.Figure at 0x1ff63ed9828>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
1
asset_niceroof
count
0
1
asset_niceroof

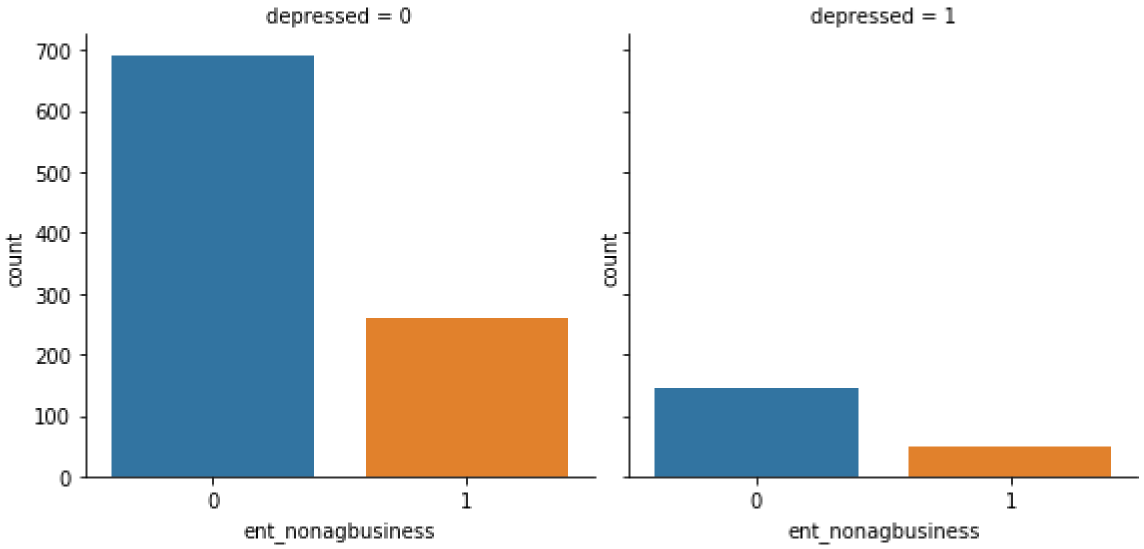
```
<matplotlib.figure.Figure at 0x1ff63feab38>
```


depressed = 0
depressed = 1
count
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
day_of_week
count
0
1
2
3
4
5
6
day_of_week

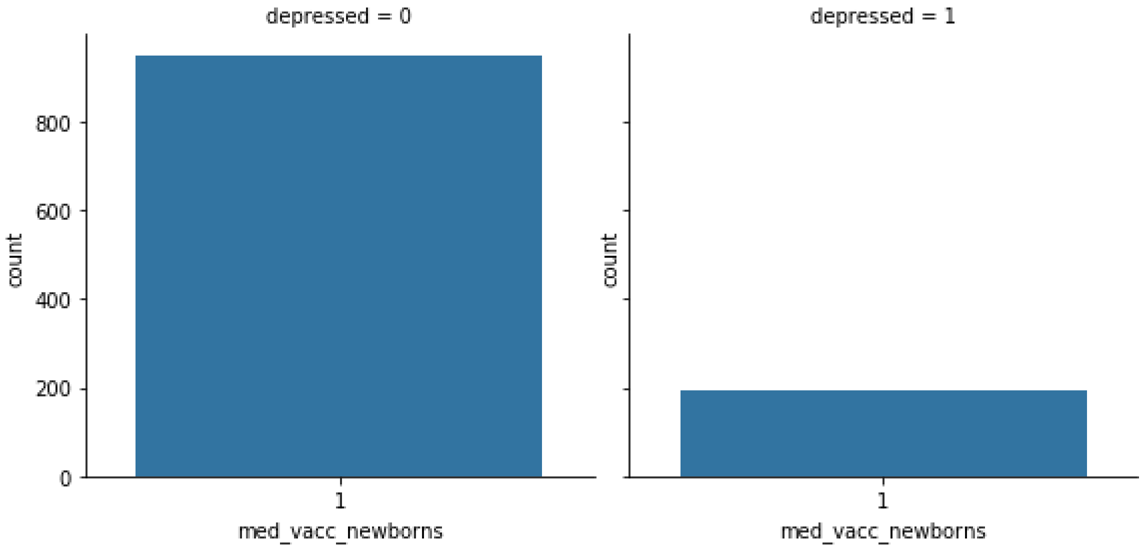
```
<matplotlib.figure.Figure at 0x1ff65682b38>
```


depressed = 0
depressed = 1
count
800
700
600
500
400
300
200
100
0
0
1
ent_business
count
0
1
ent_business

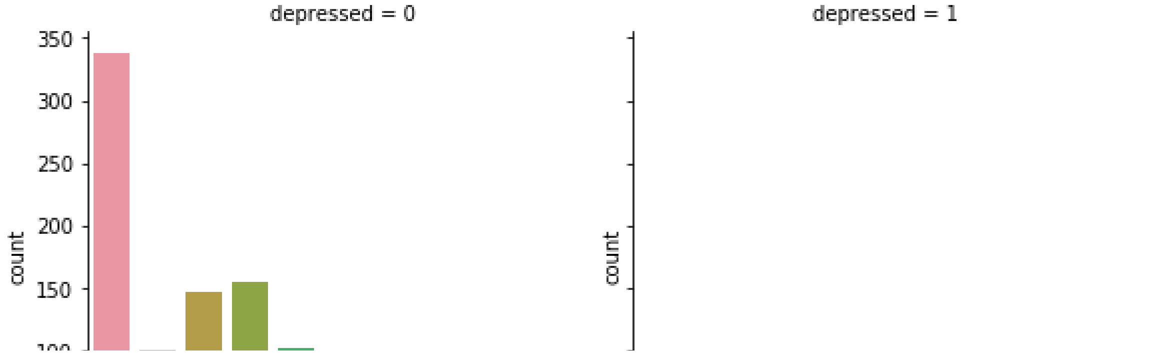
```
<matplotlib.figure.Figure at 0x1ff6589d048>
```


depressed = 0
depressed = 1
count
700
600
500
400
300
200
100
0
0
1
ent_nonagbusiness

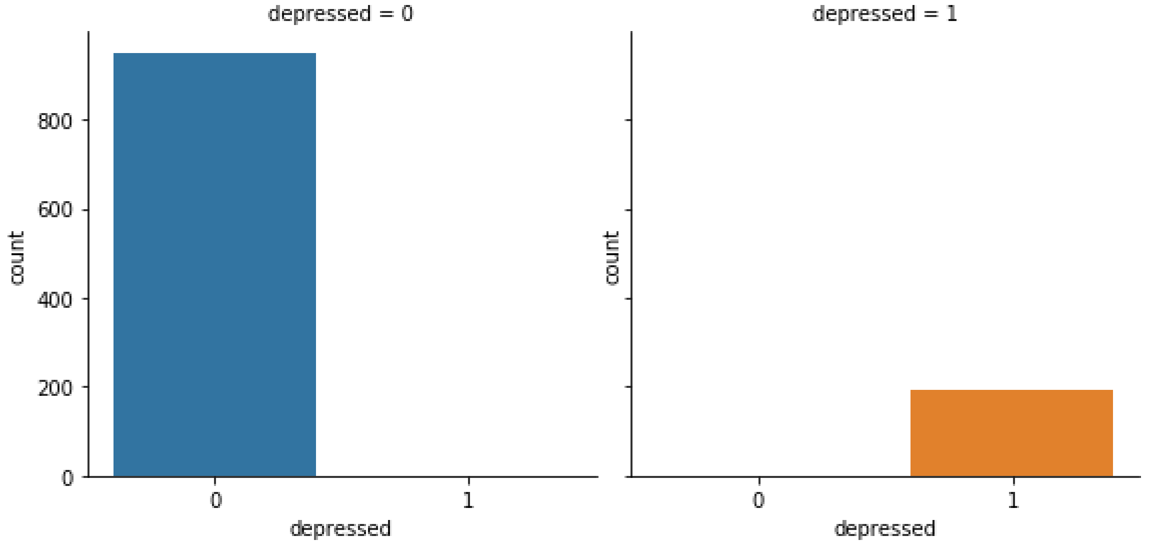
```
<matplotlib.figure.Figure at 0x1ff6419fda0>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
1
med_vacc_newborns

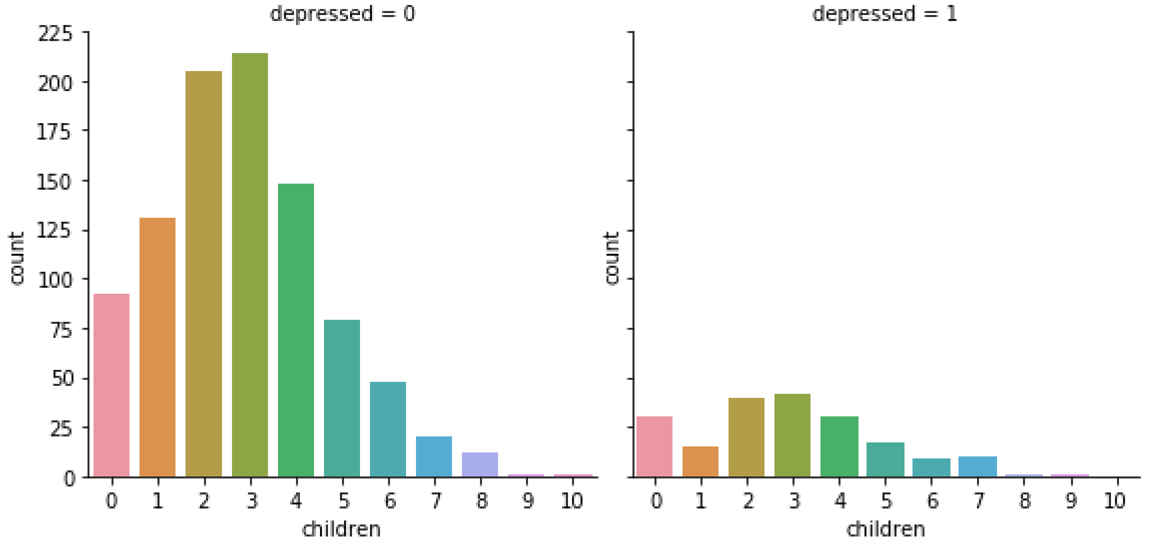
```
<matplotlib.figure.Figure at 0x1ff642dd0b8>
```


depressed = 0
depressed = 1
count
350
300
250
200
150
100
count

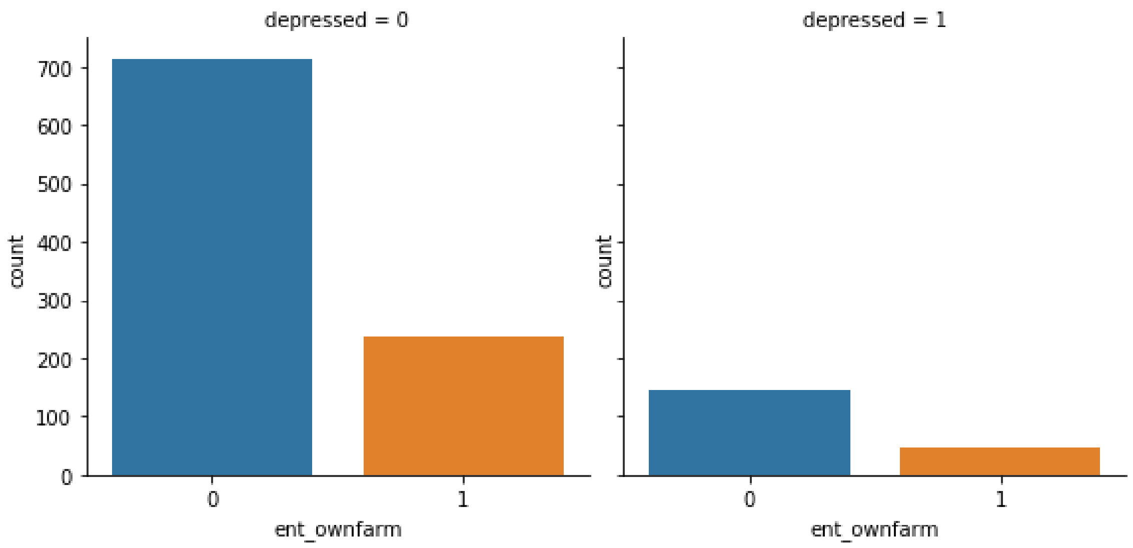
```
<matplotlib.figure.Figure at 0x1ff64191198>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
1
depressed
count
0
1
depressed

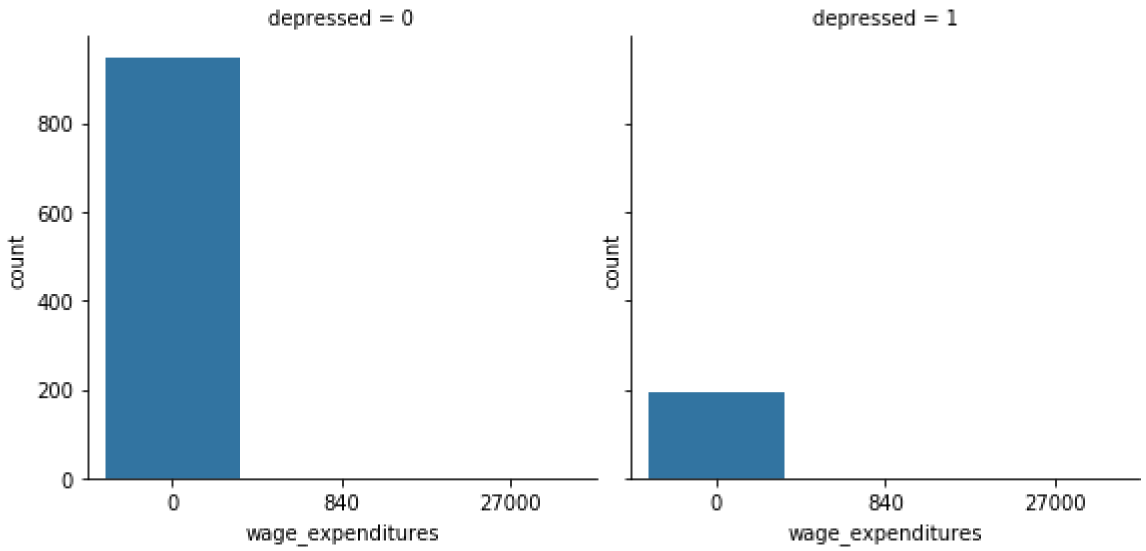
```
<matplotlib.figure.Figure at 0x1ff642d3b38>
```


depressed = 0
depressed = 1
count
225
200
175
150
125
100
75
50
25
0
0 1 2 3 4 5 6 7 8 9 10
children
count
0 1 2 3 4 5 6 7 8 9 10
children

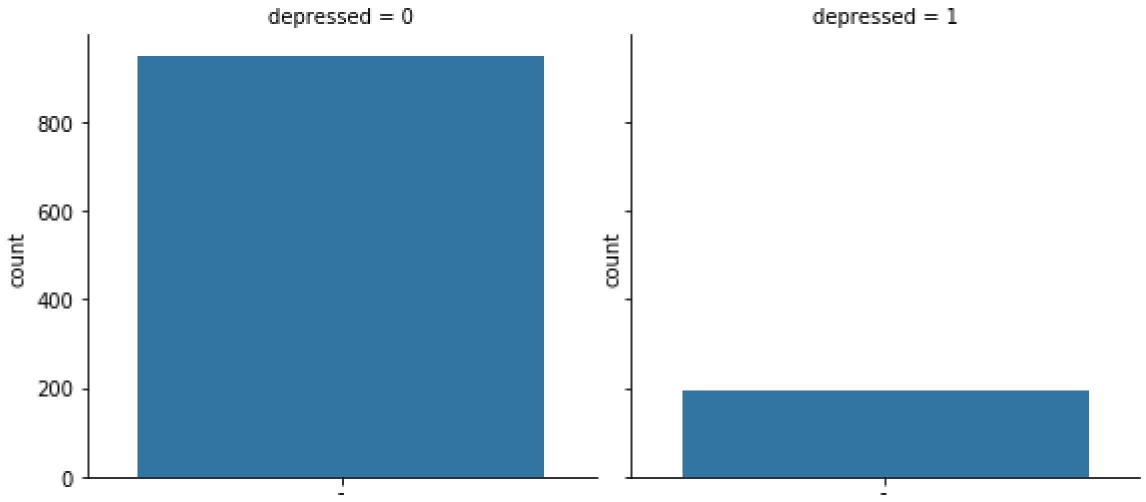
```
<matplotlib.figure.Figure at 0x1ff6423ba90>
```


depressed = 0
depressed = 1
count
700
600
500
400
300
200
100
0
0
1
ent_ownfarm
count
0
1
ent_ownfarm

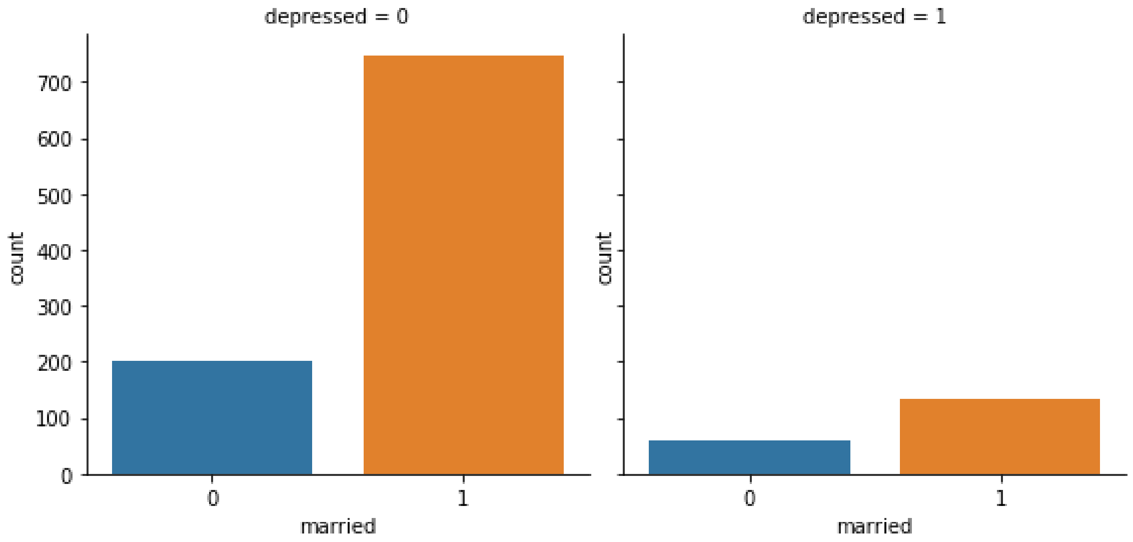
```
<matplotlib.figure.Figure at 0x1ff63f19b38>
```


depressed = 0
depressed = 1
count
800
600
400
200
0
0
840
27000
wage_expenditures
count
0
840
27000
wage_expenditures

```
<matplotlib.figure.Figure at 0x1ff641f0438>
```


depressed = 0
depressed = 1
count
count
800
600
400
200
0

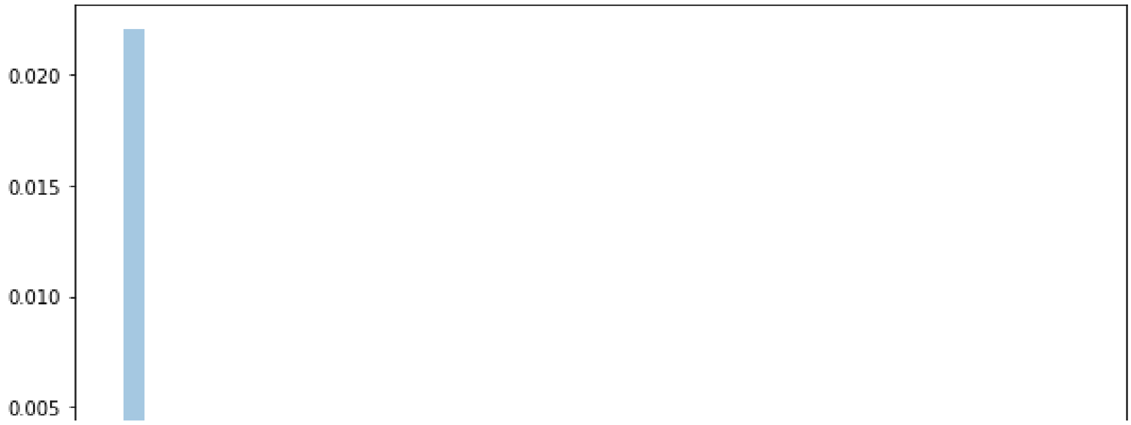
```
<matplotlib.figure.Figure at 0x1ff6427a550>
```


depressed = 0
depressed = 1
count
count
700
600
500
400
300
200
100
0
0
1
married
0
1
married

In [15]:

```
def plot_dist(data, cols):
    for col in cols:
        plt.figure(figsize=(10,5))
        sns.distplot(data[col].dropna());
        plt.show()

plot_dist(data=tr,cols=num_col)
```

executed in 34.7s, finished 21:01:56 2019-01-29



In [ ]: