nhanes_multivariate_practice

October 18, 2022

1 Practice notebook for multivariate analysis using NHANES data

This notebook will give you the opportunity to perform some multivariate analyses on your own using the NHANES study data. These analyses are similar to what was done in the week 3 NHANES case study notebook.

You can enter your code into the cells that say "enter your code here", and you can type responses to the questions into the cells that say "Type Markdown and Latex".

Note that most of the code that you will need to write below is very similar to code that appears in the case study notebook. You will need to edit code from that notebook in small ways to adapt it to the prompts below.

To get started, we will use the same module imports and read the data in the same way as we did in the case study:

```
In [10]: %matplotlib inline
         import matplotlib.pyplot as plt
         import seaborn as sns
         import pandas as pd
         import statsmodels.api as sm
         import numpy as np


         da = pd.read_csv("nhanes_2015_2016.csv")
         da.columns
```

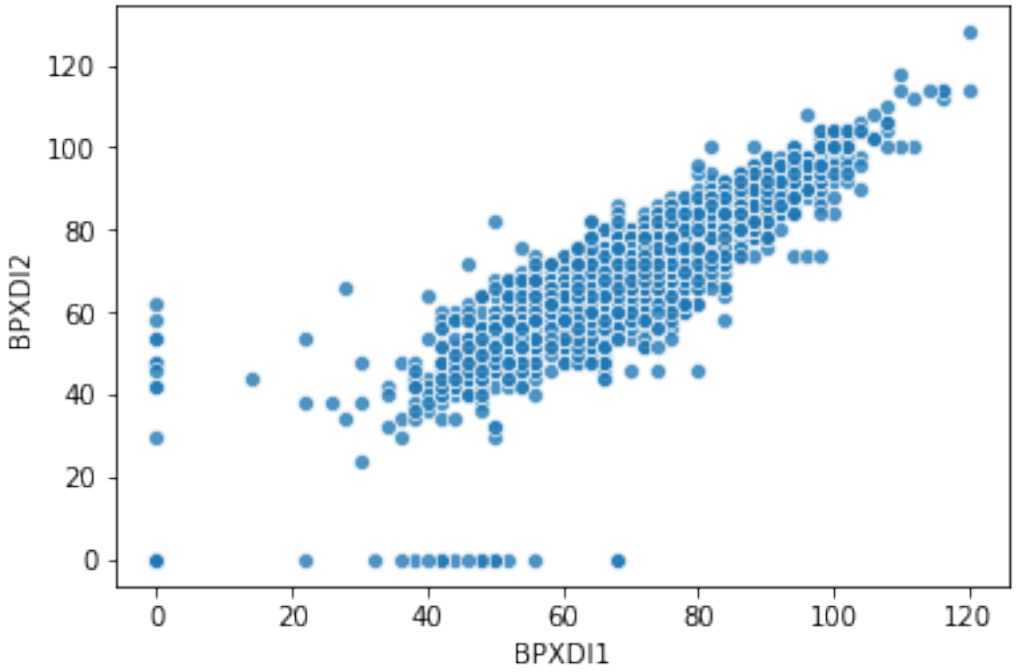
```
Out[10]: Index(['SEQN', 'ALQ101', 'ALQ110', 'ALQ130', 'SMQ020', 'RIAGENDR', 'RIDAGEYR',
                'RIDRETH1', 'DMDCITZN', 'DMDEDUC2', 'DMDMARTL', 'DMDHHSIZ', 'WTINT2YR',
                'SDMVPSU', 'SDMVSTRA', 'INDFMPIR', 'BPXSY1', 'BPXDI1', 'BPXSY2',
                'BPXDI2', 'BMXWT', 'BMXHT', 'BMXBMI', 'BMXLEG', 'BMXARML', 'BMXARMC',
                'BMXWAIST', 'HIQ210'],
               dtype='object')
```

1.1 Question 1

Make a scatterplot showing the relationship between the first and second measurements of diastolic blood pressure (BPXDI1 and BPXDI2). Also obtain the 4x4 matrix of correlation coefficients among the first two systolic and the first two diastolic blood pressure measures.

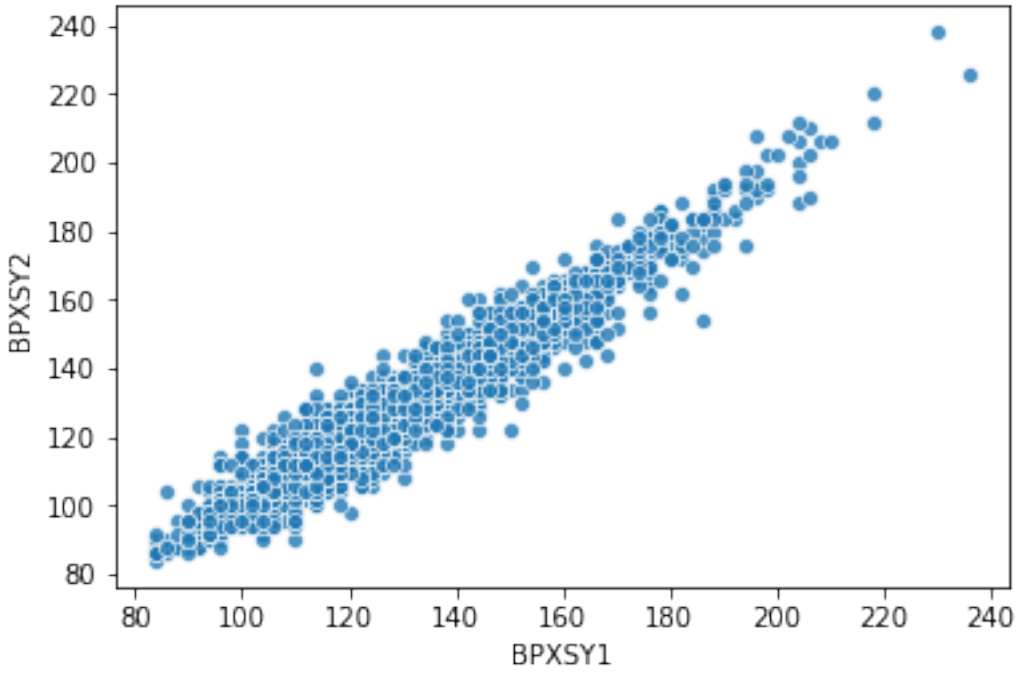
In [2]:
```
# enter your code here
sns.scatterplot(x = da["BPXDI1"], y = da["BPXDI2"], alpha = 0.8)
```

Out[2]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e1eec8b70>
```



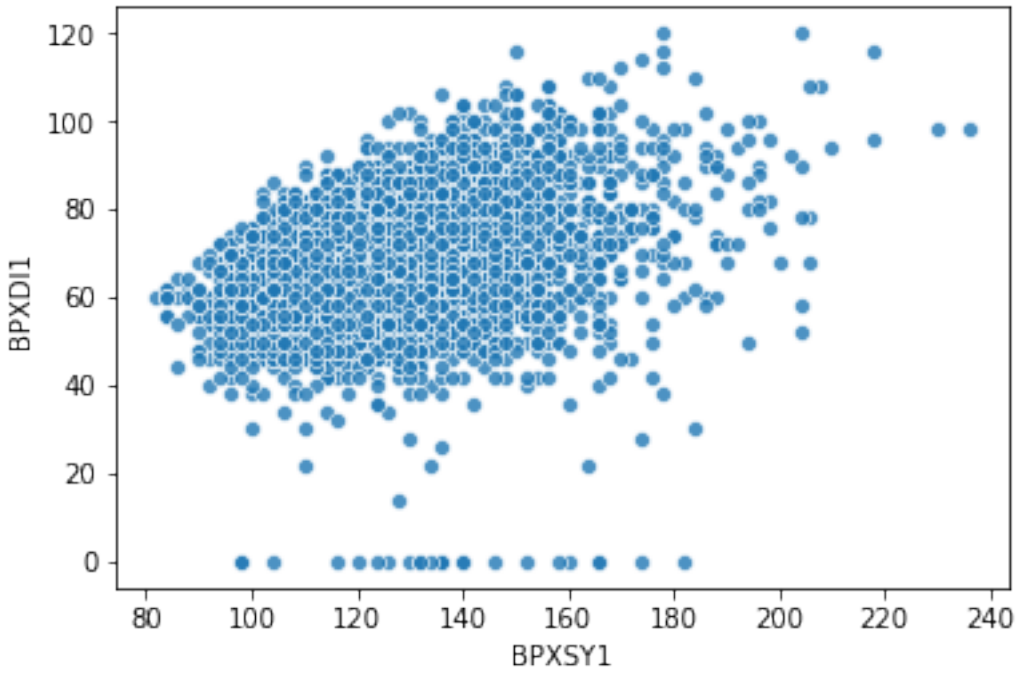
In [4]:
```
sns.scatterplot(x = da["BPXSY1"], y = da["BPXSY2"], alpha = 0.8)
```

Out[4]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e1cdb15f8>
```



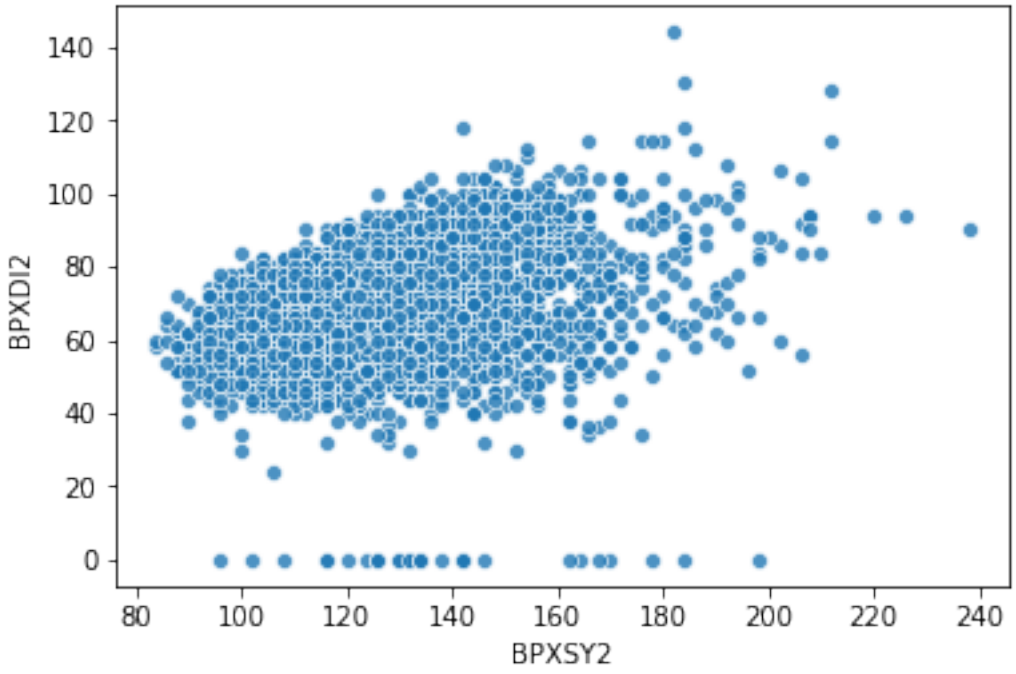

```
In [5]: sns.scatterplot(x = da["BPXSY1"], y = da["BPXDI1"], alpha = 0.8)
```

```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e1cd10588>
```



In [6]:
```
sns.scatterplot(x = da["BPXSY2"], y = da["BPXDI2"], alpha = 0.8)
```

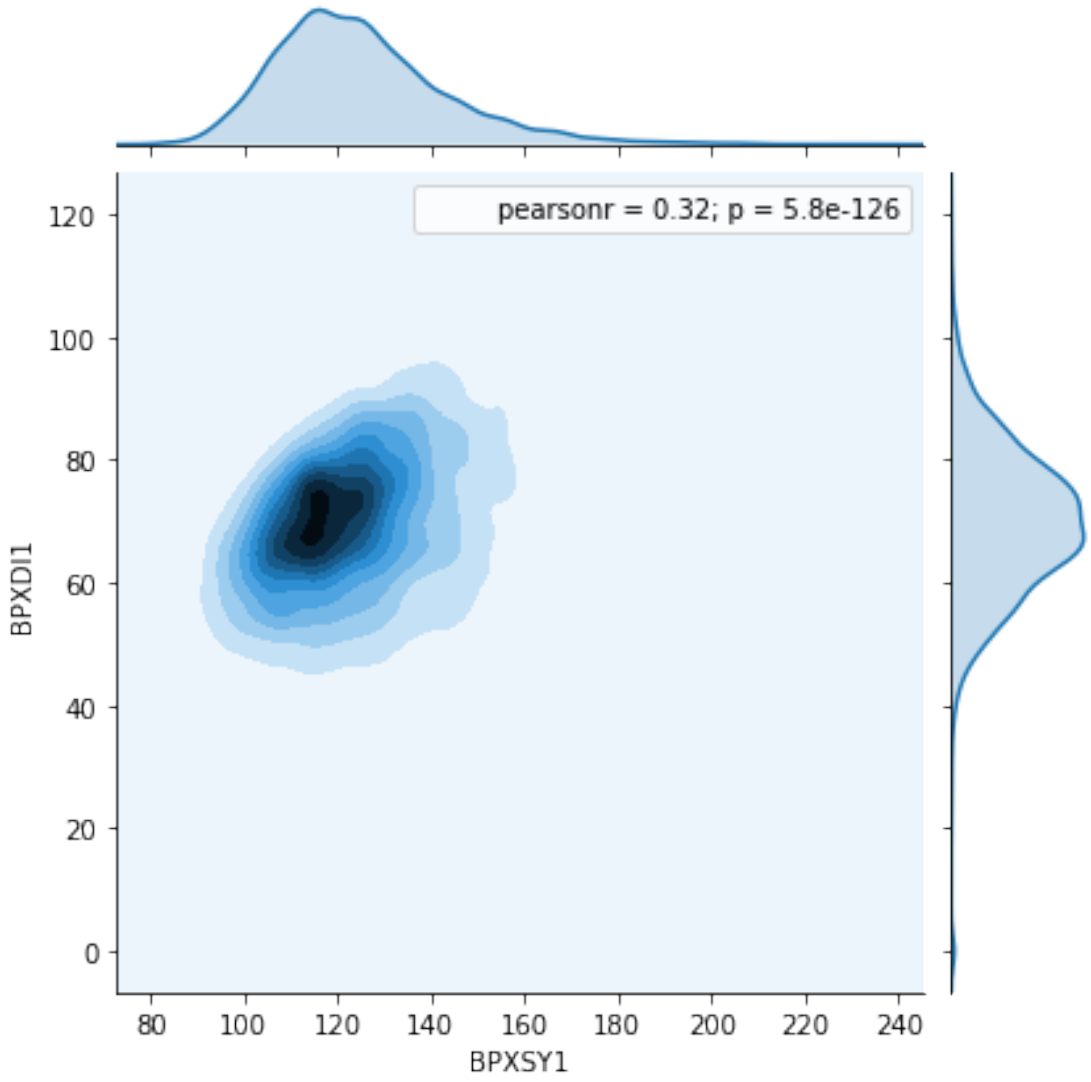
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1e1cc7e4e0>



In [15]:
```
from scipy import stats
sns.jointplot(x = "BPXSY1", y = "BPXDI1", kind ='kde', data = da).annotate(stats.pears
```

```
/opt/conda/lib/python3.6/site-packages/seaborn/axisgrid.py:1847: UserWarning: JointGrid annotat
  warnings.warn(UserWarning(msg))
```

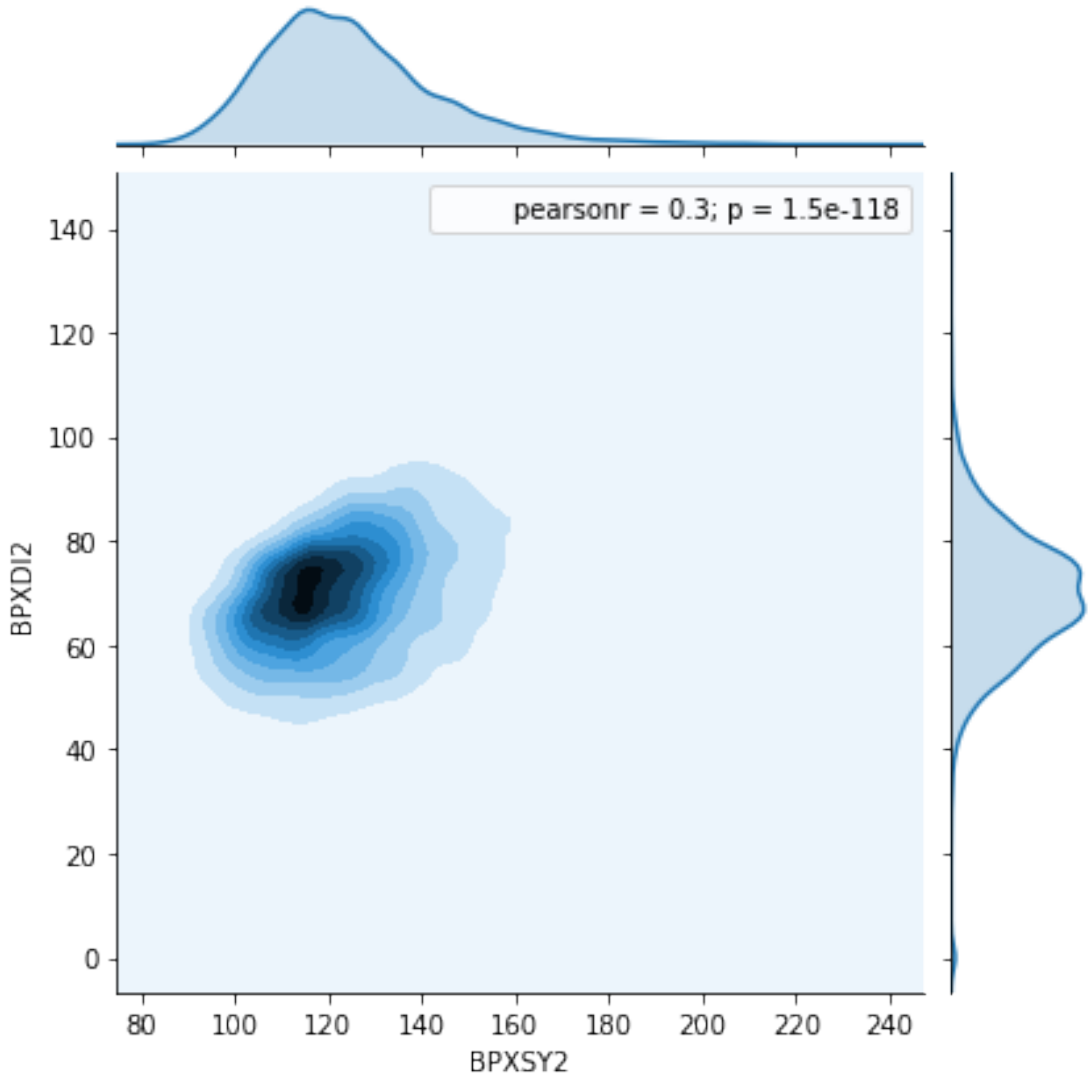
Out[15]: <seaborn.axisgrid.JointGrid at 0x7f1e1c6b7438>




```
In [16]: sns.jointplot(x = "BPXSY2", y = "BPXDI2", kind ='kde', data=da).annotate(stats.pearso

/opt/conda/lib/python3.6/site-packages/seaborn/axisgrid.py:1847: UserWarning: JointGrid annotat
  warnings.warn(UserWarning(msg))


Out[16]: <seaborn.axisgrid.JointGrid at 0x7f1e1c90f1d0>
```



Q1a. How does the correlation between repeated measurements of diastolic blood pressure relate to the correlation between repeated measurements of systolic blood pressure?

We can see that the correlation between repeated measurements of systolic blood pressure is strongly correlated than the repeated measurements of the diastolic blood pressure. It also shown that there are some potential outliers with value 0 on either measurement 1 or 2 of the diastolicc values. These need to be ignored.

Q2a. Are the second systolic and second diastolic blood pressure measure more correlated or less correlated than the first systolic and first diastolic blood pressure measure?

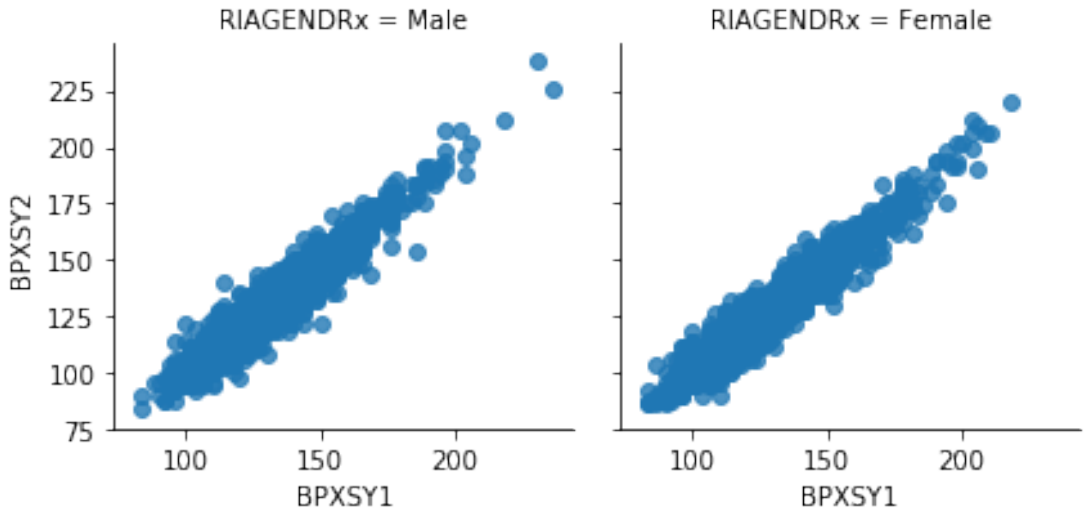
The second diastolic and systolic blood pressure readings are slightly more correlated to the first reading. The low value of 0.02 should not make much of a difference

1.2 Question 2

Construct a grid of scatterplots between the first systolic and the first diastolic blood pressure measurement. Stratify the plots by gender (rows) and by race/ethnicity groups (columns).

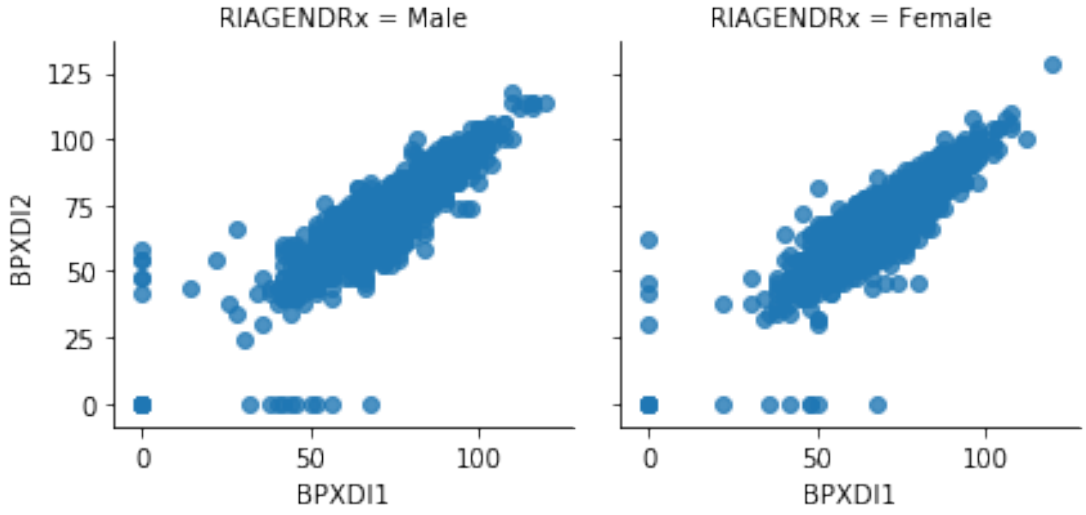
```
In [18]: # insert your code here
         da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
         sns.FacetGrid(da, col = "RIAGENDRx").map(plt.scatter, "BPXSY1", "BPXSY2", alpha = 0.8)
```

```
Out[18]: <seaborn.axisgrid.FacetGrid at 0x7f1e1cb890b8>
```



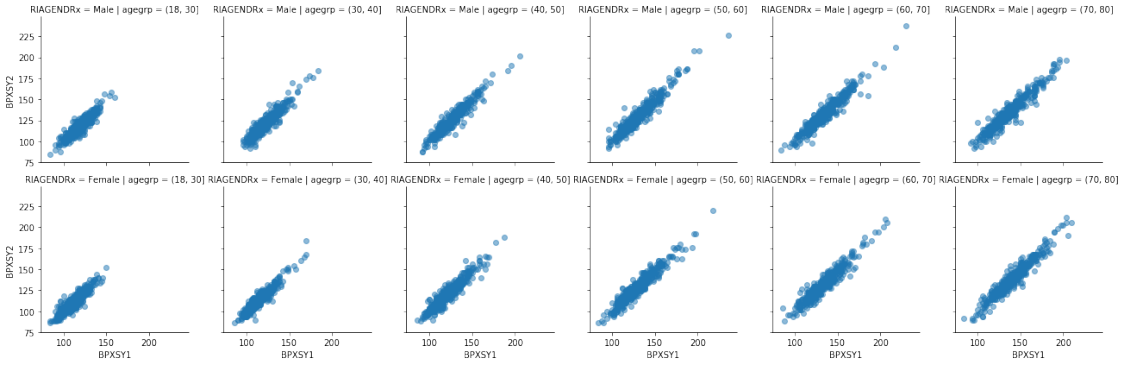

```
In [19]: da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
         sns.FacetGrid(da, col = "RIAGENDRx").map(plt.scatter, "BPXDI1", "BPXDI2", alpha = 0.8)
```

```
Out[19]: <seaborn.axisgrid.FacetGrid at 0x7f1e1c8984a8>
```



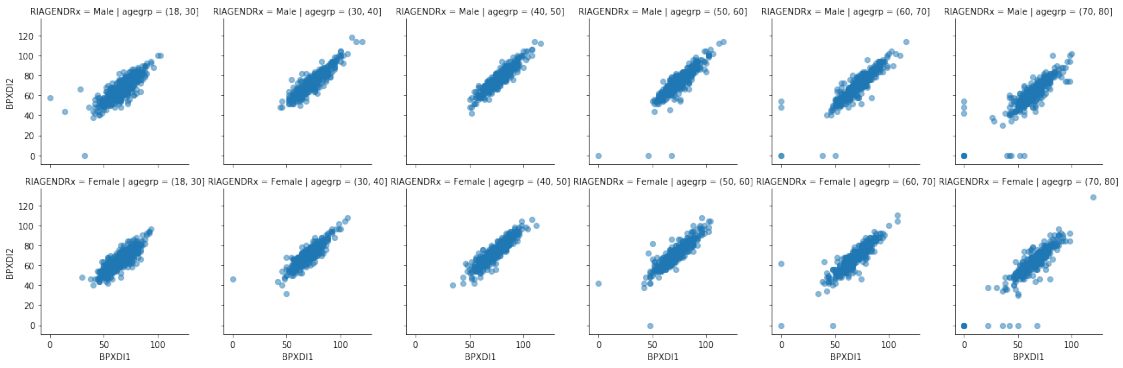
In [24]:
```
# Create age strata based on cut points
da["agegrp"] = pd.cut(da.RIDAGEYR, [18, 30, 40, 50, 60, 70, 80])
# Make the figure wider than the default (12cm wide by 5cm tall)
plt.figure(figsize=(12, 10))
_=sns.FacetGrid(da, col = "agegrp", row = "RIAGENDRx").map(plt.scatter, "BPXSY1", "BPX
```

```
<Figure size 864x720 with 0 Axes>
```



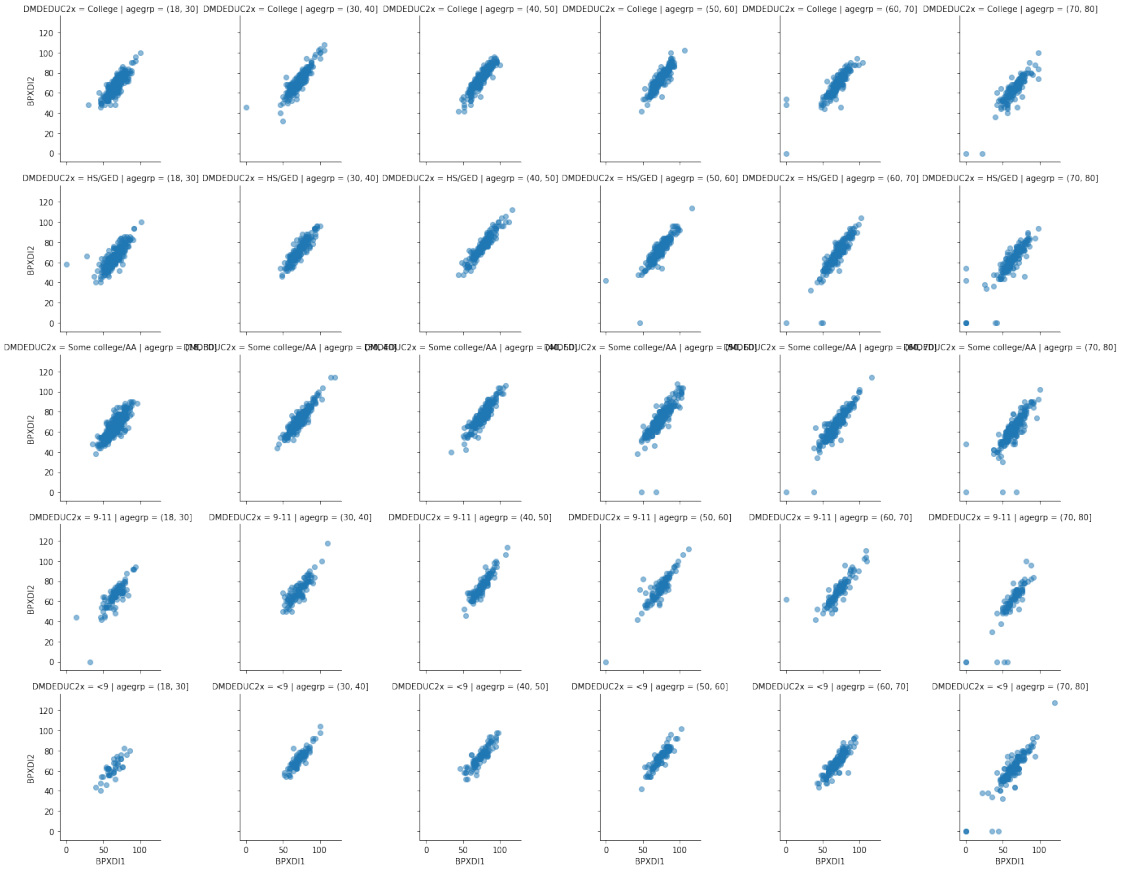
In [25]:
```
plt.figure(figsize=(12, 10))
_=sns.FacetGrid(da, col = "agegrp", row = "RIAGENDRx").map(plt.scatter, "BPXDI1", "BPX
```

```
<Figure size 864x720 with 0 Axes>
```

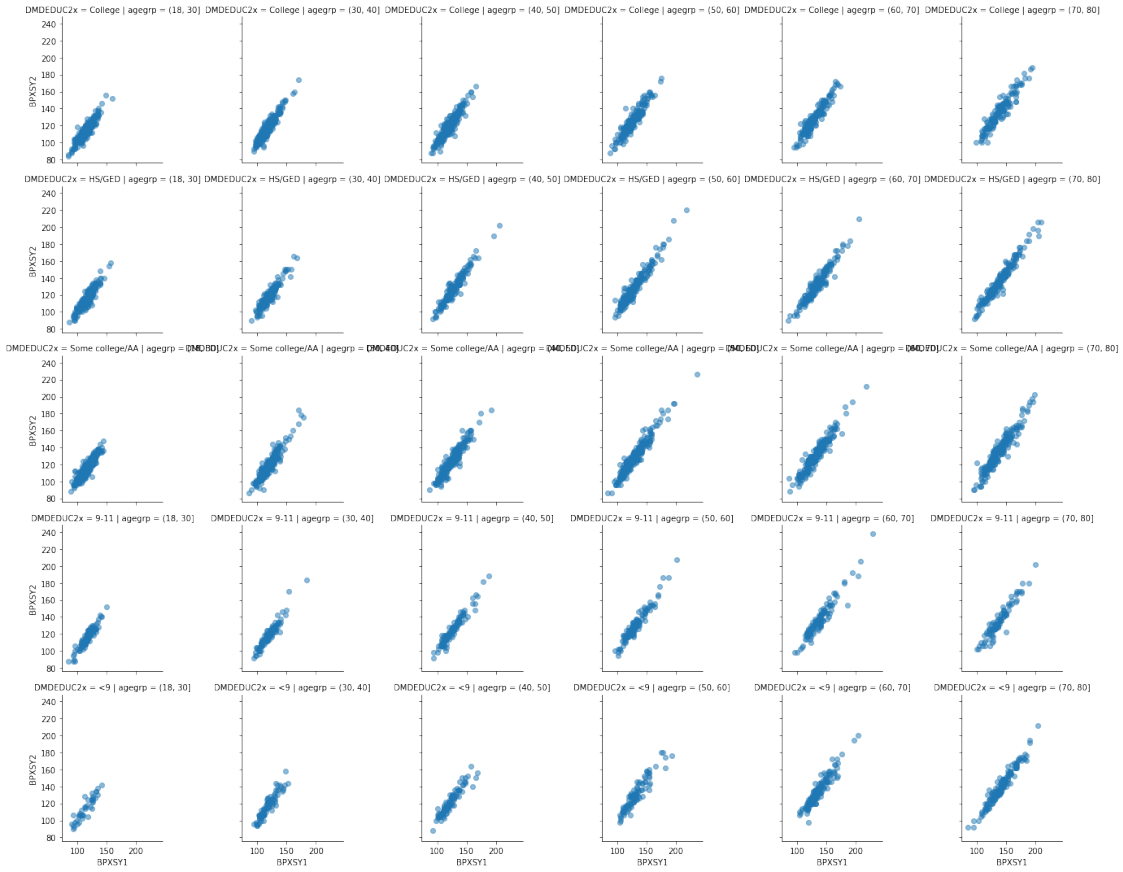


In [27]:
```
da["DMDEDUC2x"] = da.DMDEDUC2.replace({1: "<9", 2: "9-11", 3: "HS/GED", 4: "Some colle
                                       7: "Refused", 9: "Don't know"})
da["DMDMARTLx"] = da.DMDEDUC2.replace({1: "Married", 2: "Widowed", 3: "Divorced", 4:
```

```
                                                6: "Living w/partner", 77: "Refused"})
df2 = da.loc[(da.DMDEDUC2x != "Don't know") & (da.DMDMARTLx != "Refused"), :]
_=sns.FacetGrid(df2, col="agegrp", row="DMDEDUC2x").map(plt.scatter,"BPXDI1", "BPXDI2
```



In [28]: `_=sns.FacetGrid(df2, col="agegrp", row="DMDEDUC2x").map(plt.scatter,"BPXSY1", "BPXSY2`

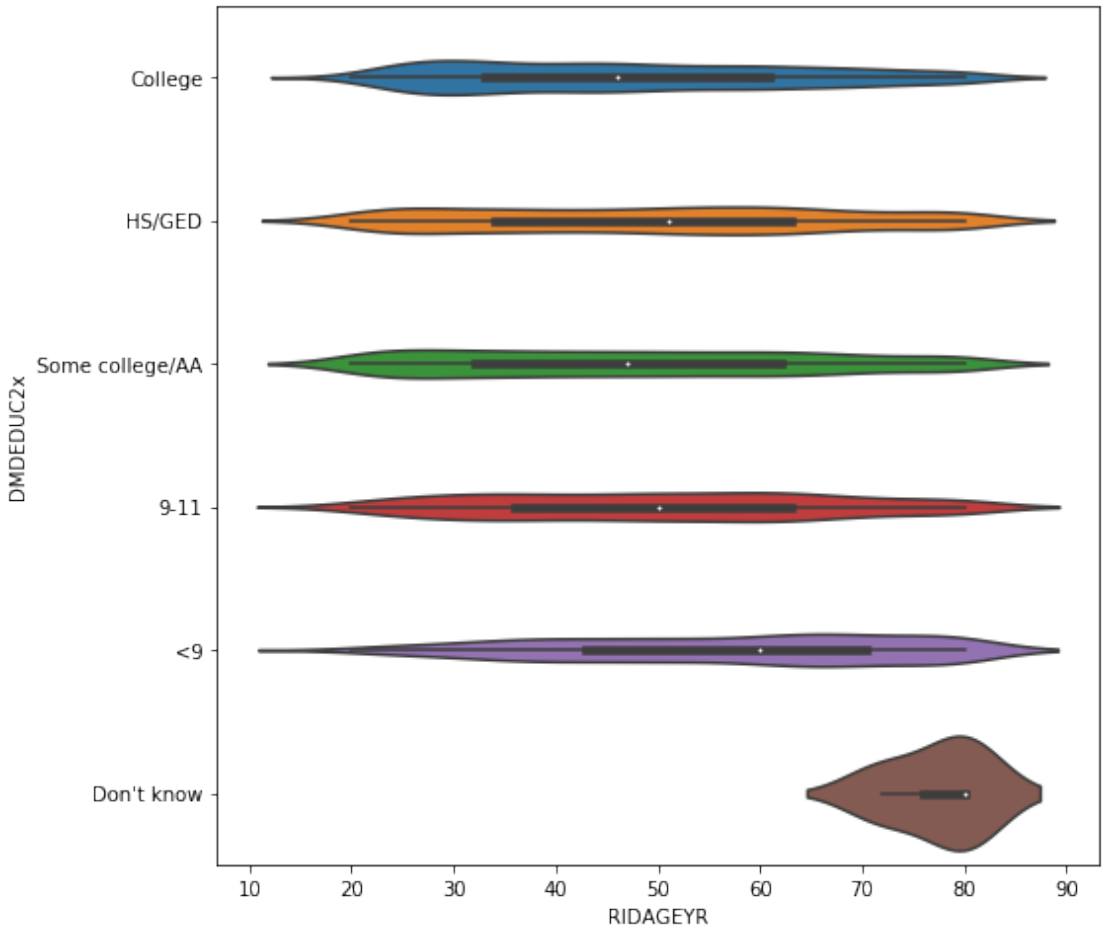


Q3a. Comment on the extent to which these two blood pressure variables are correlated to different degrees in different demographic subgroups.

1.3 Question 3

Use "violin plots" to compare the distributions of ages within groups defined by gender and educational attainment.

```
In [29]: # insert your code here
         plt.figure(figsize=(8, 8))
         a = sns.violinplot(da.RIDAGEYR, da.DMDEDUC2x)
```



Q4a. Comment on any evident differences among the age distributions in the different demographic groups.

1.4 Question 4

Use violin plots to compare the distributions of BMI within a series of 10-year age bands. Also stratify these plots by gender.

```
In [ ]: # insert your code here
```

Q5a. Comment on the trends in BMI across the demographic groups.

1.5 Question 5

Construct a frequency table for the joint distribution of ethnicity groups (RIDRETH1) and health-insurance status (HIQ210). Normalize the results so that the values within each ethnic group are proportions that sum to 1.

```
In [ ]: # insert your code here
```

Q6a. Which ethnic group has the highest rate of being uninsured in the past year?