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 NH 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 a 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 [1]:
```
%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[1]:
```
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')
```

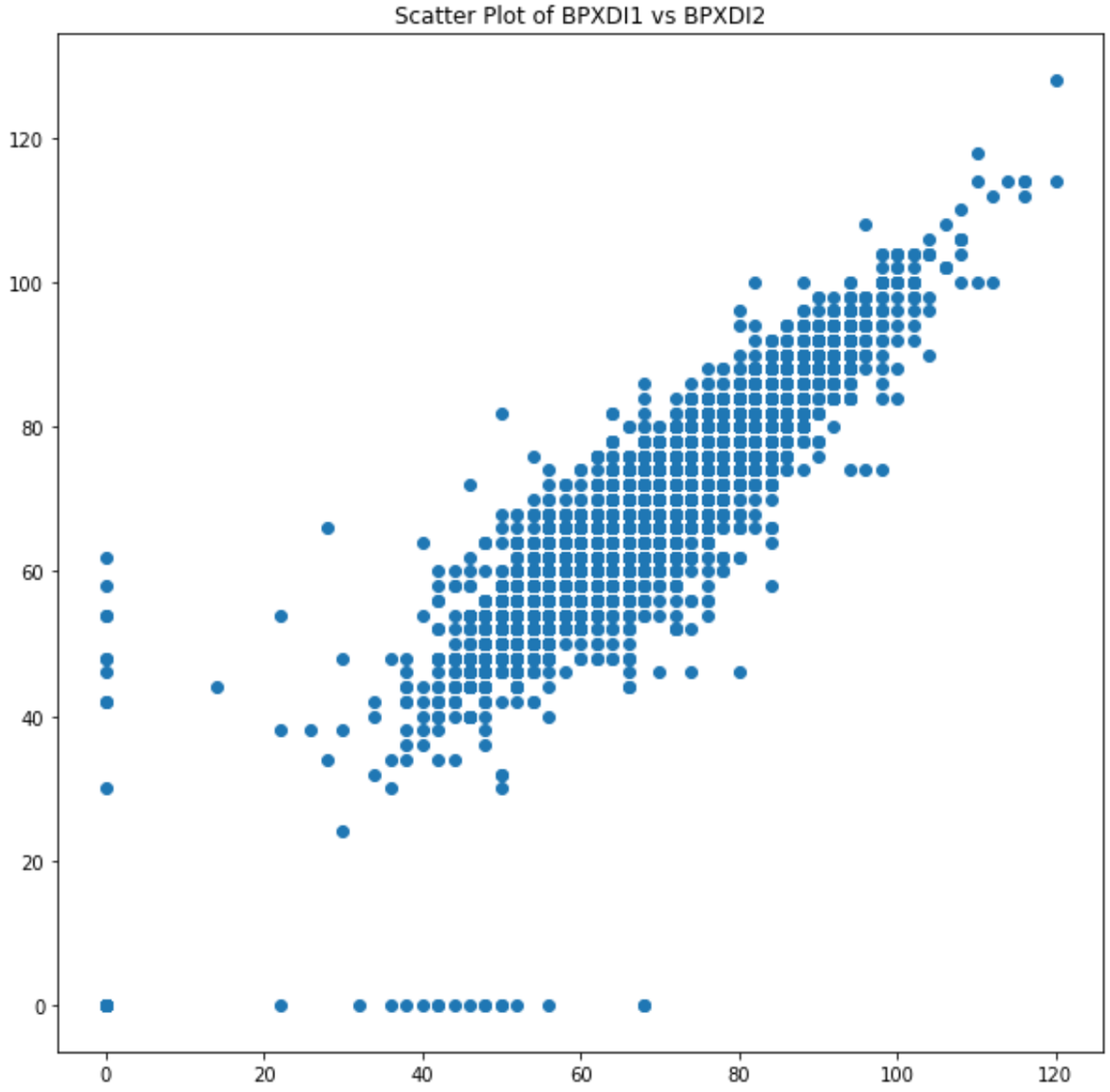
Question 1

Make a scatterplot showing the relationship between the first and second measurements of diastolic blood pressure (BPXDI1and BPXDI2). Also obtain the 4x4 matrix of correlation coefficier among the first two systolic and the first two diastolic blood pressure measures.

In [2]:

```
# enter your code here
plt.figure(figsize=(10,10))
plt.scatter(x = da.BPXDI1, y = da.BPXDI2)
plt.title("Scatter Plot of BPXDI1 vs BPXDI2")
plt.show()

df = da[['BPXSY1', 'BPXSY2', 'BPXDI1', 'BPXDI2']] #create new data frame with needed columns
df.corr()
```


Scatter Plot of BPXDI1 vs BPXDI2
120
100
80
60
40
20
0
0
20
40
60
80
100
120

Out[2]:

	BPXSY1	BPXSY2	BPXDI1	BPXDI2
BPXSY1	1.000000	0.962287	0.316531	0.277681
BPXSY2	0.962287	1.000000	0.329843	0.303847
BPXDI1	0.316531	0.329843	1.000000	0.884722
BPXDI2	0.277681	0.303847	0.884722	1.000000

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

==> Slightly less correleated

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?

==> First slightly more correlated

Question 2

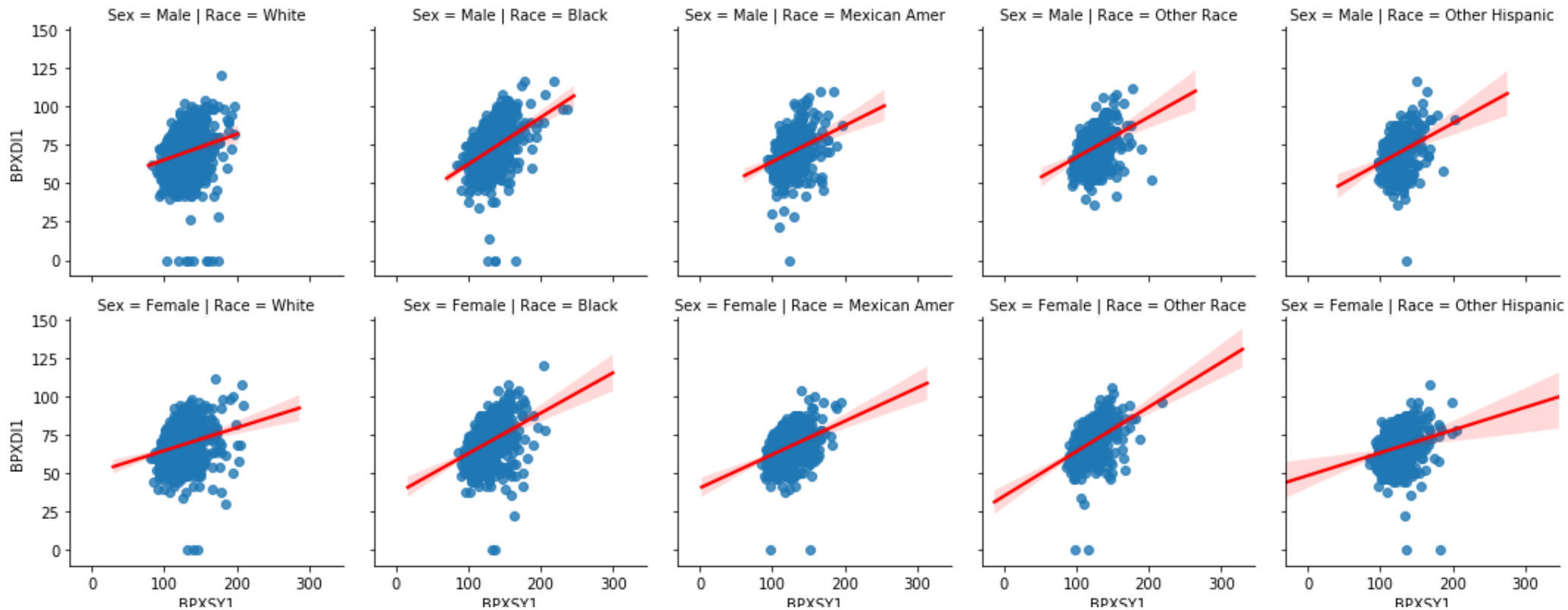
Construct a grid of scatterplots between the first systolic and the first diastolic blood pressure measurement. Stratify the plots by gender (rows) and by race/ethnicity groups (columns).

In [3]:
```
# insert your code here
df = da[['BPXSY1', 'BPXDI1', 'BPXSY2', 'BPXDI2', 'RIAGENDR', 'RIDRETH1']]
df["RIAGENDR"] = df.RIAGENDR.replace({1: 'Male', 2: 'Female'})
# recode the educational variable
df["RIDRETH1"] = df.RIDRETH1.replace({1: "Mexican Amer", 2: "Other Hispanic",
                                      3: "White", 4: "Black",
                                      5: "Other Race"})
df.rename(columns={"RIAGENDR": "Sex", "RIDRETH1": "Race"}, inplace = True)

plt.figure(figsize=(20,10))
graph = sns.FacetGrid(df, 'Sex', 'Race')
graph.map(sns.regplot, 'BPXSY1', 'BPXDI1', line_kws={"color": "red"})
```

Out[3]:
```
<seaborn.axisgrid.FacetGrid at 0x7f64a9d703c8>

<Figure size 1440x720 with 0 Axes>
```



Q2a. Comment on the extent to which these two blood pressure variables are correlated to different degrees in different demographic subgroups.

All are to some degree correlated; Black Males most correlated; White Females least

Question 3

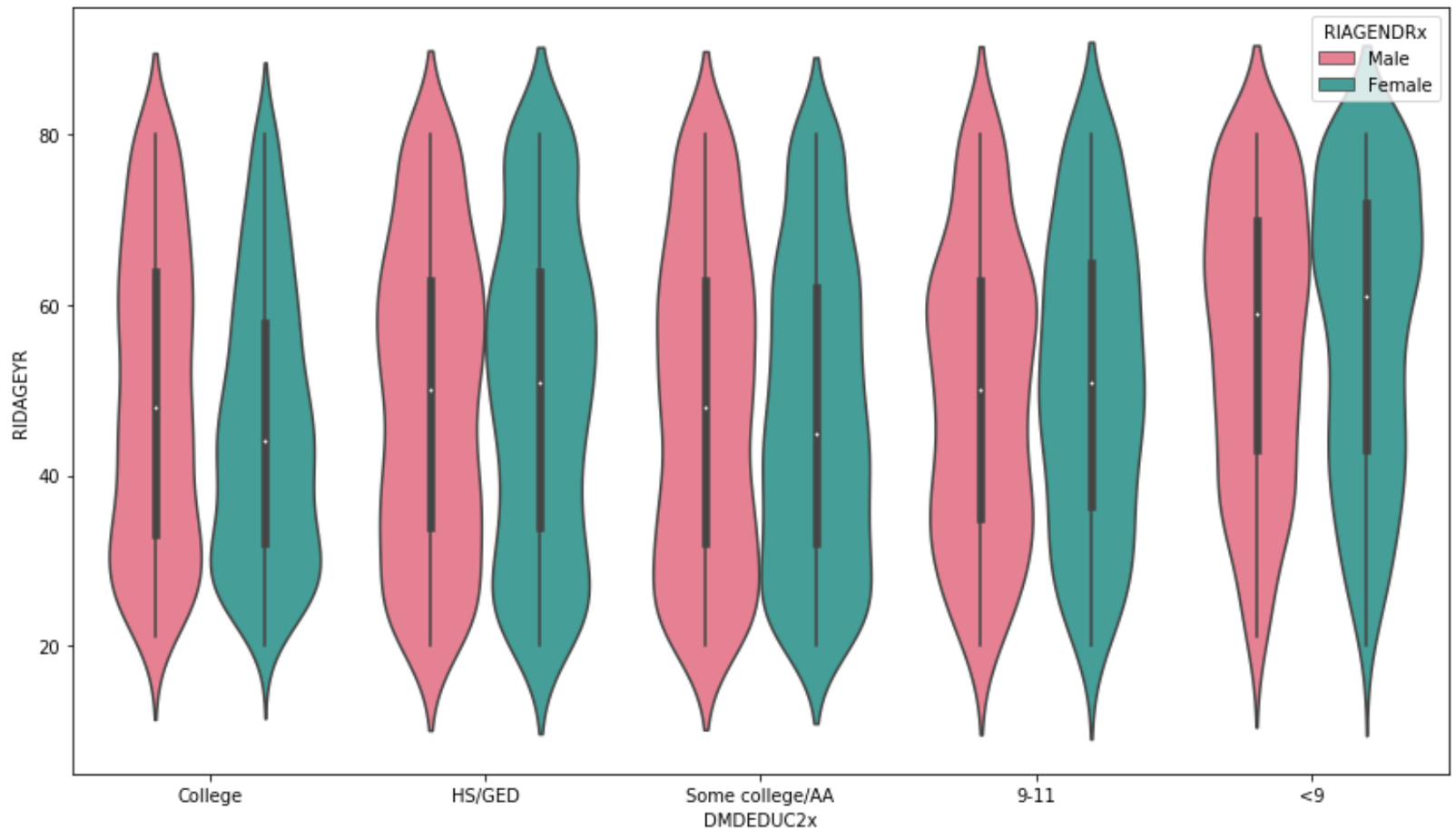
Use "violin plots" to compare the distributions of ages within groups defined by gender and educational attainment.

In [10]:
```
# insert your code here
plt.figure(figsize=(14, 8))
da["DMDEDUC2x"] = da.DMDEDUC2.replace({1: "<9", 2: "9-11", 3: "HS/GED", 4: "Some college/AA", 5: "College", 7: "Refused", 9: "Don'
w"})

da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
da["AGEGRP"] = pd.cut(da.RIDAGEYR, [18, 30, 40, 50, 60, 70, 80])

df = da[da.DMDEDUC2 <= 5]
sns.violinplot("DMDEDUC2x", "RIDAGEYR", hue = "RIAGENDRx" , data = df, palette="husl")
```

Out[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6517b95588>
```



Q3a. Comment on any evident differences among the age distributions in the different demographic groups.

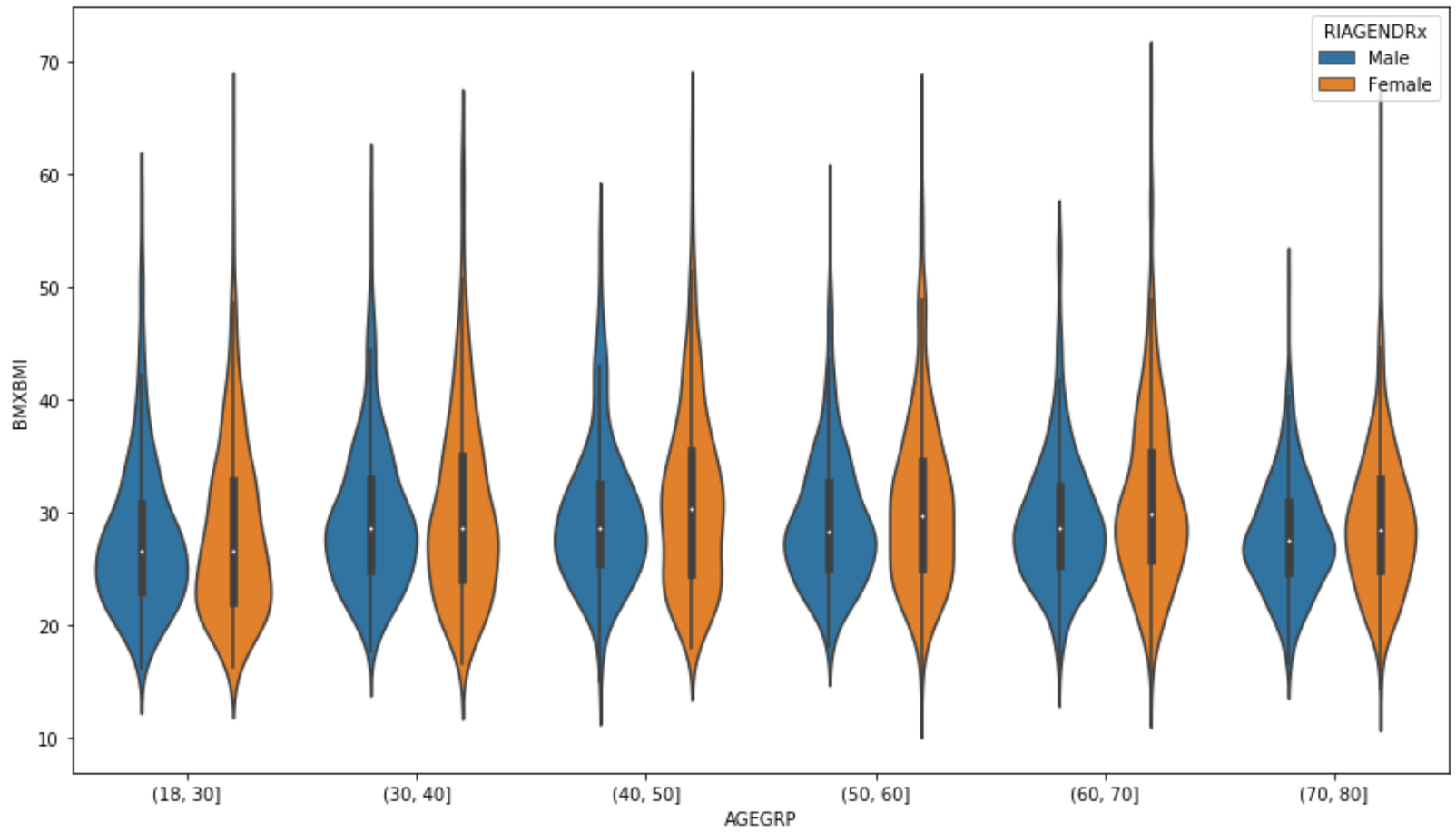
Less educated tend to be older; college-educated younger

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 [9]:
```
# insert your code here
plt.figure(figsize=(14, 8))
sns.violinplot("AGEGRP", "BMXBMI", hue="RIAGENDRx", data=df)
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6517ba85f8>
```



Q4a. Comment on the trends in BMI across the demographic groups.

These are some fat violins

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 a proportions that sum to 1.

In [11]:
```
# insert your code here
df = da[['HIQ210', 'RIDRETH1']]
df["RIDRETH1"] = df.RIDRETH1.replace({1: "Mexican Amer", 2: "Other Hispanic",
                                      3: "White", 4: "Black",
                                      5: "Other Race"})
df["HIQ210"] = df.HIQ210.replace({1: "Yes", 2: "No", 9:"Unknown"})

print("Uninsured in past year by Race")
dx = df.groupby(["RIDRETH1"])["HIQ210"].value_counts()
dx = dx.unstack()
dx = dx.apply(lambda x: x/x.sum(), axis=1)
print(dx.to_string(float_format="%.3f"), '\n')
```

```
Uninsured in past year by Race
HIQ210            No  Unknown   Yes
RIDRETH1
Black          0.890    0.001 0.109
Mexican Amer   0.858    0.004 0.138
Other Hispanic 0.871      NaN 0.129
Other Race     0.916    0.003 0.082
White          0.931    0.001 0.067
```

Q5a. Which ethnic group has the highest rate of being uninsured in the past year?

Mexican Americans