(b)

In [142]:

```
corr = df2.iloc[:,2:].corr()
# corr.to_csv("corr.csv", index = False)
idx = list(corr.columns)
# find the minimun correlation
for row in range(corr.shape[0]): # df is the DataFrame
    for col in range(corr.shape[1]):
        if corr.iloc[row,col] == sorted(corr.min())[0]:
            print(sorted(corr.min())[0])
            print(idx[row], idx[col])
        if corr.iloc[row,col] == sorted(corr.min())[2]:
            print(sorted(corr.min())[2])
            print(idx[row], idx[col])

corr = corr.replace(1, 0)
# find the maximun correlation
for row in range(corr.shape[0]): # df is the DataFrame
    for col in range(corr.shape[1]):
        if corr.iloc[row,col] == sorted(corr.max(), reverse = True)[0]:
            print(sorted(corr.max(), reverse = True)[0])
            print(idx[row], idx[col])
        if corr.iloc[row,col] == sorted(corr.max(), reverse = True)[2]:
            print(sorted(corr.max(), reverse = True)[2])
            print(idx[row], idx[col])
```

```
0.5684269349479398
DRG Charges 269 DRG Charges 371
0.5835260020031648
DRG Charges 315 DRG Charges 460
0.5684269349479398
DRG Charges 371 DRG Charges 269
0.5835260020031648
DRG Charges 460 DRG Charges 315
0.9612236108383365
DRG Charges 292 DRG Charges 293
0.9612236108383365
DRG Charges 293 DRG Charges 292
0.977653142783413
DRG Charges 481 DRG Charges 482
0.977653142783413
DRG Charges 482 DRG Charges 481
```

For high positive association:
Correlation between DRG Charges 292 and DRG Charges 293 is 0.9612236108383365.
Correlation between DRG Charges 481 and DRG Charges 482 is 0.977653142783413.
From the first two plot, we can see that pairs of DRG Charges with high associations are almost linear, which is indicated by correlations.

For low positive association:
Correlation between DRG Charges 269 and DRG Charges 371 is 0.5684269349479398.
Correlation between DRG Charges 315 and DRG Charges 460 is 0.5835260020031648.
From the last two plot, we can see that these two pairs of DRG Charges's relationship is not very obvious. By using ggplot to add a smooth on it, we obtain a curve. While for the first two plot, what we get is almost a straight line. The correlations for these two pairs are much lower than the fisrt two pair and it support the observation.

3.2 Boxplots and T-tests

(a)

Acccording to the GDP of different states, select CA, TX, GA, PA, IN, ME to exhibit differences in their hospital charges.

In [10]:

```
import pandas as pd

pd.unique(df2['Provider State'])
states = ["CA", "TX","GA","PA","IN","ME"]
df_6states = df2[df2['Provider State'].isin(states)]
df_6states.to_csv("DRG_6states.csv", index = False)
```

Preprocess data by pandas, which makes it easier to manipulate in ggplot2.

In [51]:

```
df_boxplot1 = df_6states[['DRG Charges 190','Provider State']]
df_boxplot1.dropna()
df_boxplot1.to_csv("df_boxplot1.csv", index = False)

df_boxplot2 = df_6states[['DRG Charges 392','Provider State']]
df_boxplot2.dropna()
df_boxplot2.to_csv("df_boxplot2.csv", index = False)

df_boxplot3 = df_6states[['DRG Charges 871','Provider State']]
df_boxplot3.dropna()
df_boxplot3.to_csv("df_boxplot3.csv", index = False)
```