# baby_Clustering

April 24, 2018

In [85]:
```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import seaborn as sns
sns.set(style="whitegrid", color_codes=True)
from pylab import *
import warnings
warnings.filterwarnings(action='once')
```

In [86]:
```
df = pd.read_csv('babyNames.csv')
# from https://data.world/data-society/most-popular-baby-names-in-nyc
# birth year, gender, ethnicity, name, count, rank
```

In [87]:
```
df.head()
```

Out[87]:

| | BRTH_YR | GNDR | ETHCTY | NM | CNT | RNK |
|---|---|---|---|---|---|---|
| 0 | 2011 | FEMALE | HISPANIC | GERALDINE | 13 | 75 |
| 1 | 2011 | FEMALE | HISPANIC | GIA | 21 | 67 |
| 2 | 2011 | FEMALE | HISPANIC | GIANNA | 49 | 42 |
| 3 | 2011 | FEMALE | HISPANIC | GISELLE | 38 | 51 |
| 4 | 2011 | FEMALE | HISPANIC | GRACE | 36 | 53 |

In [88]:
```
df.isnull().any()
```

Out[88]:
```
BRTH_YR    False
GNDR       False
ETHCTY     False
NM         False
CNT        False
RNK        False
dtype: bool
```

In [89]:
```
le_gender = LabelEncoder()
le_ethnicity = LabelEncoder()
le_name = LabelEncoder()
```

```
le_gender.fit(df['GNDR'])
df['GNDR'] = le_gender.transform(df['GNDR'])
le_ethnicity.fit(df['ETHCTY'])
df['ETHCTY'] = le_ethnicity.transform(df['ETHCTY'])
le_name.fit(df['NM'])
df['NM'] = le_name.transform(df['NM'])
df.head()
```

Out[89]:

| | BRTH_YR | GNDR | ETHCTY | NM | CNT | RNK |
|---|---|---|---|---|---|---|
| 0 | 2011 | 0 | 4 | 1019 | 13 | 75 |
| 1 | 2011 | 0 | 4 | 1021 | 21 | 67 |
| 2 | 2011 | 0 | 4 | 1023 | 49 | 42 |
| 3 | 2011 | 0 | 4 | 1028 | 38 | 51 |
| 4 | 2011 | 0 | 4 | 1036 | 36 | 53 |

In [90]:
```
cluster_num = 3
km = KMeans(n_clusters=cluster_num)
```

In [91]:
```
km.fit(df)
```

Out[91]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=3, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

In [92]:
```
km.cluster_centers_
```

Out[92]:
```
array([[2.01179080e+03, 4.96673190e-01, 3.85342466e+00, 1.42481389e+03,
        3.59900196e+01, 5.64340509e+01],
       [2.01186511e+03, 4.83763530e-01, 3.83825978e+00, 4.32985012e+02,
        3.51586178e+01, 5.62793505e+01],
       [2.01201136e+03, 4.84436759e-01, 4.02445652e+00, 2.35508547e+03,
        3.19466403e+01, 5.83777174e+01]])
```

In [93]:
```
df['predicted_cluster'] = km.predict(df)
```

In [94]:
```
zero = df['predicted_cluster'] == 0
one = df['predicted_cluster'] == 1
two = df['predicted_cluster'] == 2
df = df.drop('predicted_cluster',axis=1)
```

In [95]:
```
print('Cluster One Length: {}'.format(len(df[zero])))
print('Cluster Two Length: {}'.format(len(df[one])))
print('Cluster Three Length: {}'.format(len(df[two])))
```

```
Cluster One Length: 5112
Cluster Two Length: 4801
Cluster Three Length: 4049
```

In [96]:
```
cs1 = df[zero]
cs2 = df[one]
cs3 = df[two]
```

### 0.1 all custer contains all birth year

```
In [97]: print(cs1['BRTH_YR'].unique(),cs2['BRTH_YR'].unique(),cs3['BRTH_YR'].unique())
```

```
[2011 2012 2013 2014] [2011 2012 2013 2014] [2011 2012 2013 2014]
```

### 0.2 all custer contains both gender

```
In [98]: print(le_gender.inverse_transform(cs1['GNDR'].unique()),
                le_gender.inverse_transform(cs1['GNDR'].unique()),
                le_gender.inverse_transform(cs1['GNDR'].unique()))
```

```
['FEMALE' 'MALE'] ['FEMALE' 'MALE'] ['FEMALE' 'MALE']
```

```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/sklearn/preprocessing/label.py:
  if diff:
```

```
In [99]: print(le_ethnicity.inverse_transform(cs1['ETHCTY'].unique()))
         print(le_ethnicity.inverse_transform(cs1['ETHCTY'].unique()))
         print(le_ethnicity.inverse_transform(cs1['ETHCTY'].unique()))
```

```
['HISPANIC' 'WHITE NON HISPANIC' 'ASIAN AND PACIFIC ISLANDER'
 'BLACK NON HISPANIC' 'ASIAN AND PACI' 'BLACK NON HISP' 'WHITE NON HISP']
['HISPANIC' 'WHITE NON HISPANIC' 'ASIAN AND PACIFIC ISLANDER'
 'BLACK NON HISPANIC' 'ASIAN AND PACI' 'BLACK NON HISP' 'WHITE NON HISP']
['HISPANIC' 'WHITE NON HISPANIC' 'ASIAN AND PACIFIC ISLANDER'
 'BLACK NON HISPANIC' 'ASIAN AND PACI' 'BLACK NON HISP' 'WHITE NON HISP']
```

```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/sklearn/preprocessing/label.py:
  if diff:
```

### 0.3 unique names in diffrent clusters

```
In [100]: cs1_name = le_name.inverse_transform(cs1['NM'].unique())
          cs2_name = le_name.inverse_transform(cs2['NM'].unique())
          cs3_name = le_name.inverse_transform(cs3['NM'].unique())
          print(len(cs1_name),len(cs2_name),len(cs3_name))
```

```
961 929 921
```

```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/sklearn/preprocessing/label.py:
  if diff:
```

### 0.4 from following data it look like cluster 1 has more high ranking names, 2 less and 3 more lesser. It can be checked through unique

```
In [101]: cs1['CNT'].max(),cs1['CNT'].min(),cs1['RNK'].max(),cs1['RNK'].min()
```

```
Out[101]: (426, 10, 102, 1)
```

```
In [102]: cs2['CNT'].max(),cs2['CNT'].min(),cs2['RNK'].max(),cs2['RNK'].min()
```

```
Out[102]: (304, 10, 102, 1)
```

```
In [103]: cs3['CNT'].max(),cs3['CNT'].min(),cs3['RNK'].max(),cs3['RNK'].min()
```

```
Out[103]: (291, 10, 102, 1)
```

```
In [104]: all_names = le_name.inverse_transform(df['NM'].unique())
          print('Total {} unique names '.format(len(all_names)))
```

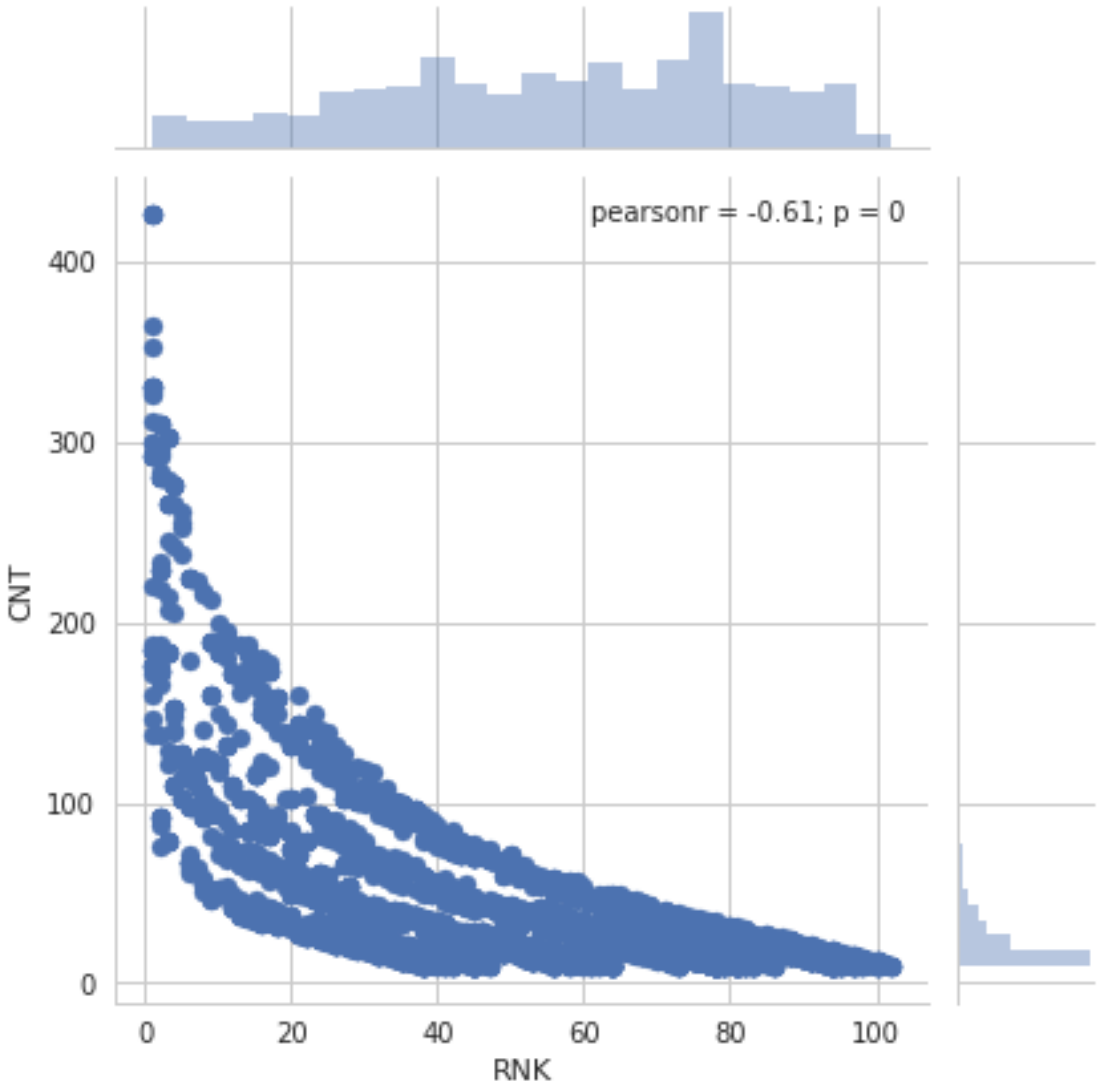
```
Total 2811 unique names
```

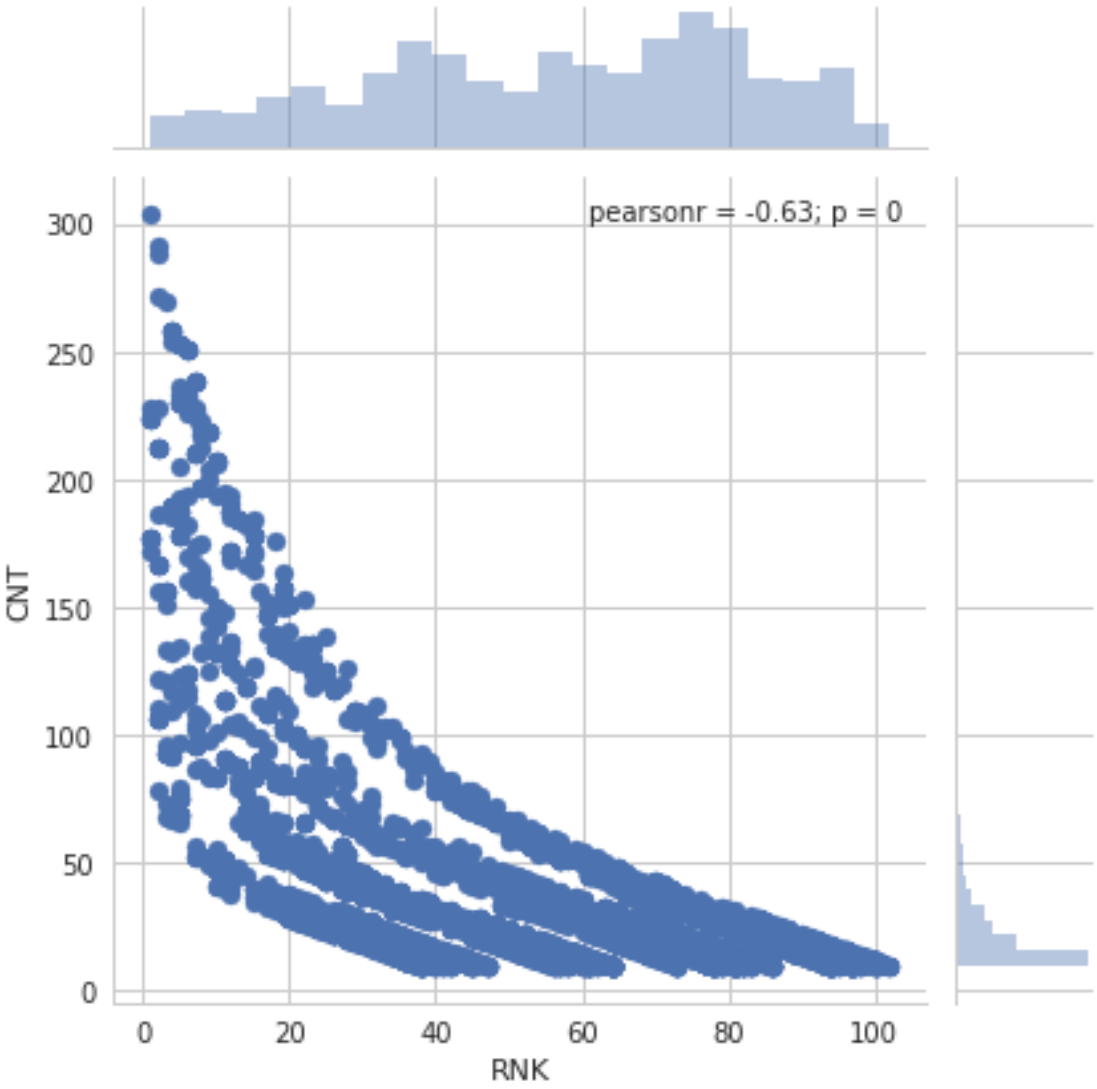
```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/sklearn/preprocessing/label.py:
  if diff:
```

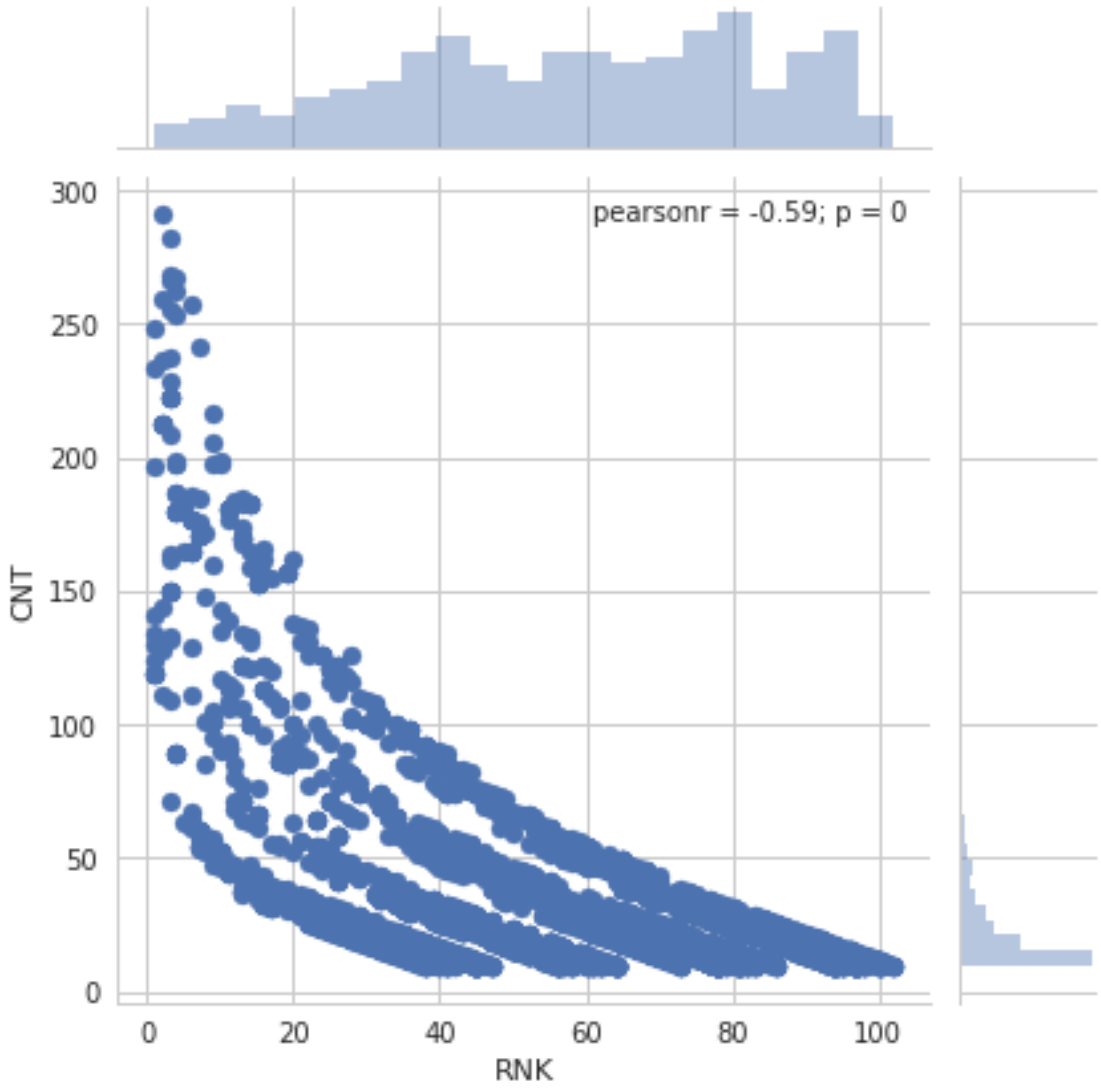
```
In [105]: # sns.pairplot(df)
          # sns.jointplot(x="GNDR",y="CNT",data=df)
          sns.jointplot(x="RNK",y="CNT",data=cs1)
          sns.jointplot(x="RNK",y="CNT",data=cs2)
          sns.jointplot(x="RNK",y="CNT",data=cs3)
```

```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462:
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
Out[105]: <seaborn.axisgrid.JointGrid at 0x7f7d0f72a588>
```


pearsonr = -0.61; p = 0
CNT
RNK
0
100
200
300
400
0
20
40
60
80
100

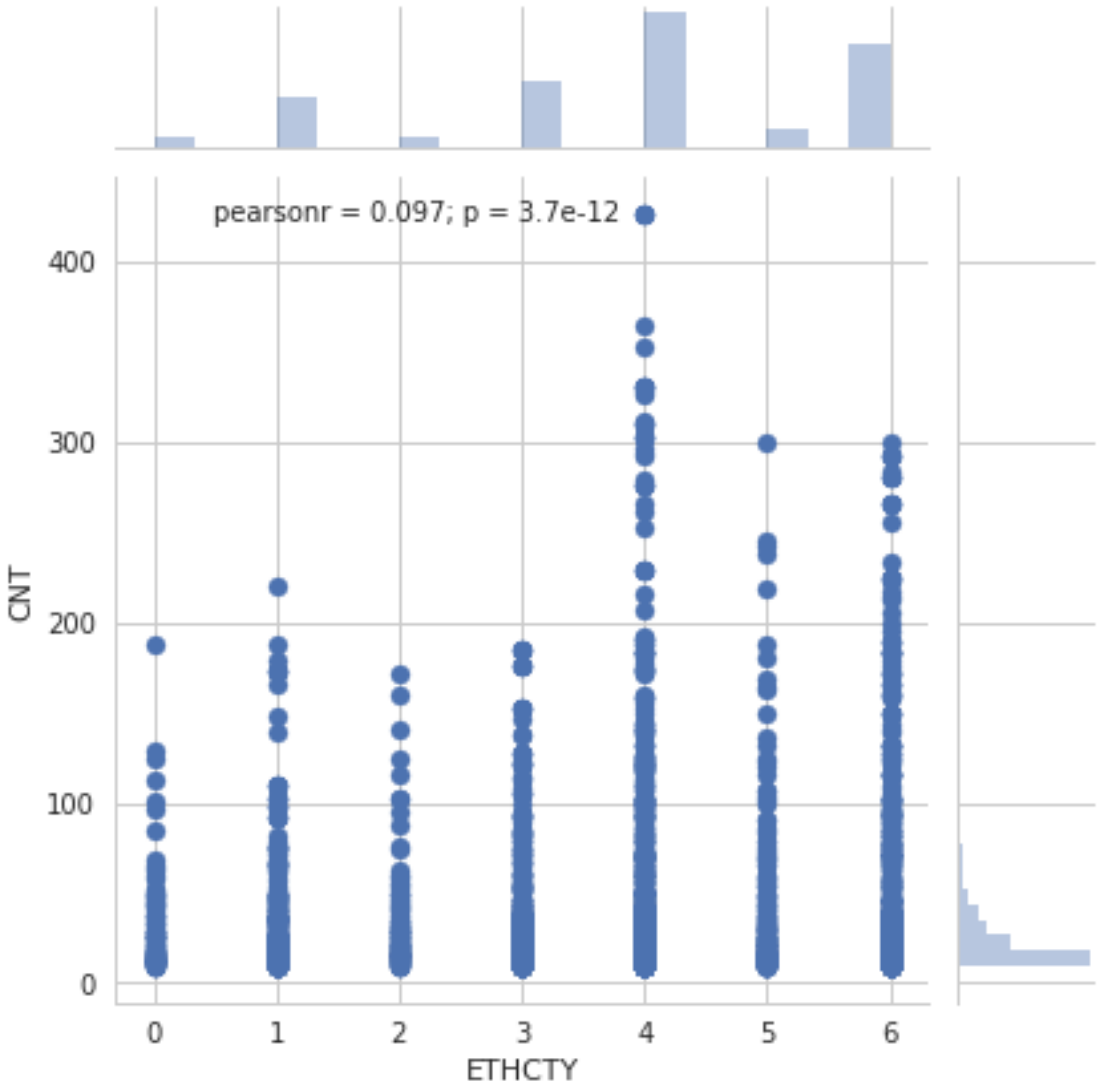

pearsonr = -0.63; p = 0
CNT
RNK
0
50
100
150
200
250
300
0
20
40
60
80
100

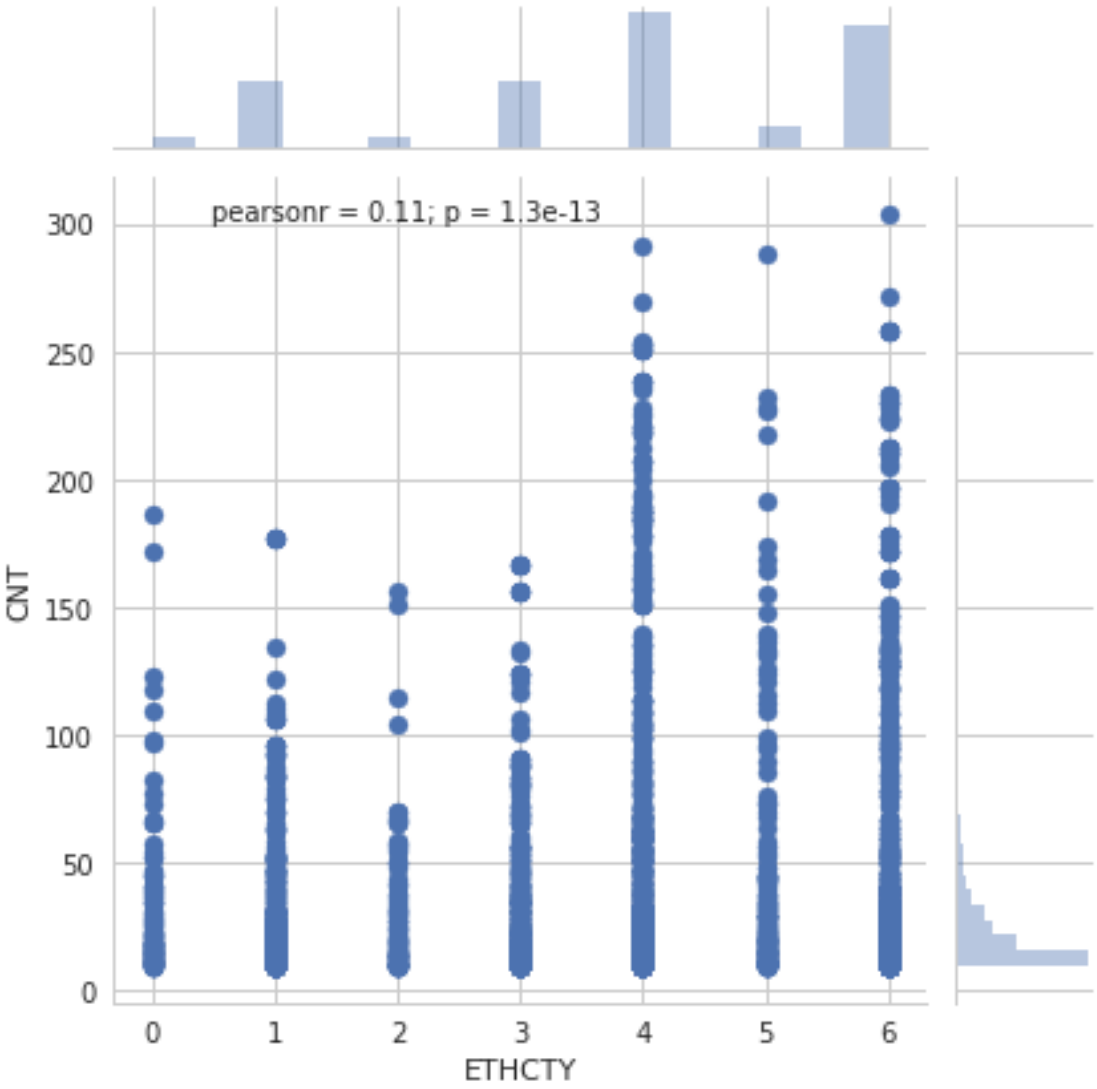


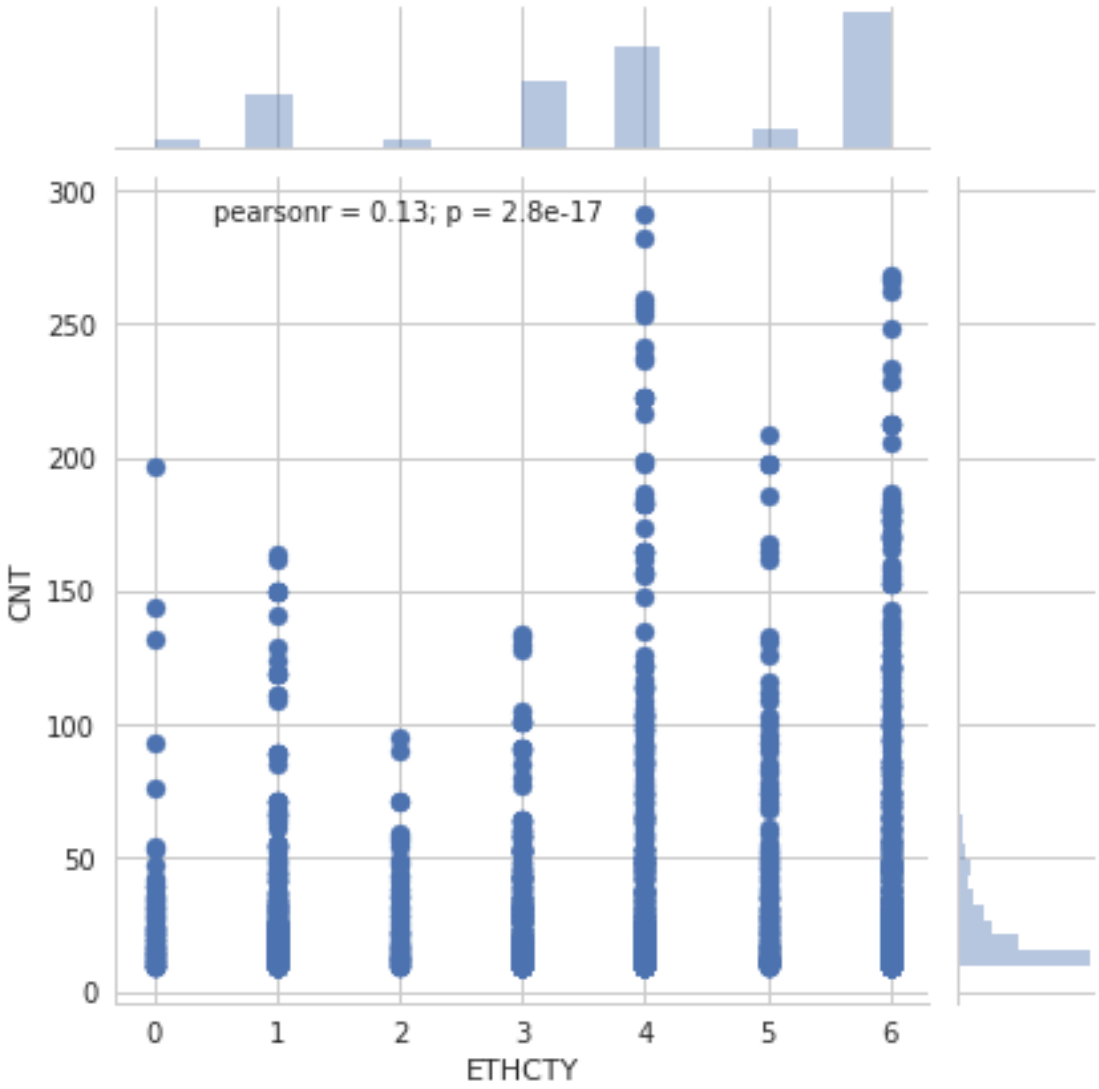

In [106]:
```
# sns.jointplot(x="ETHCTY",y="CNT",data=df)#
sns.jointplot(x="ETHCTY",y="CNT",data=cs1)
sns.jointplot(x="ETHCTY",y="CNT",data=cs2)
sns.jointplot(x="ETHCTY",y="CNT",data=cs3)
```

```
/home/multiplexer/anaconda3/envs/py3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462:
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

Out[106]: <seaborn.axisgrid.JointGrid at 0x7f7d0f86e630>


pearsonr = 0.097; p = 3.7e-12
CNT
ETHCTY
0
100
200
300
400
0
1
2
3
4
5
6


pearsonr = 0.11; p = 1.3e-13
CNT
ETHCTY
0
50
100
150
200
250
300
0
1
2
3
4
5
6


pearsonr = 0.13; p = 2.8e-17
CNT
ETHCTY
0
50
100
150
200
250
300
0
1
2
3
4
5
6