In [4]:
```
from notebooks.utils import *
```

In [5]:
```
import biovec
import numpy as np
import pandas as pd
from itertools import chain
```

In [6]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [7]:
```
from sklearn.svm import SVC
from sklearn.metrics import make_scorer
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, f1_score, roc_auc_score, mat
thews_corrcoef
from sklearn.model_selection import train_test_split
```

In [8]:
```
from sklearn.manifold import TSNE
from sklearn.decomposition import PCA
```

In [9]:
```
avp_ic50 = pd.read_csv("data/raw/AVP-IC50Pred_train.csv")
ha_avp = pd.read_csv("data/raw/HA_AVP.csv")

df = pd.concat([avp_ic50[['Sequence', 'MIC']], ha_avp], axis=0).drop_dup
licates(['Sequence']).reset_index(drop=True)
df = sequence_filtering(df)
```

In [10]:
```
df['lessthan5'] = df['MIC'].apply(lambda mic: 1 if mic <= 5 else 0)
```

In [11]:
```
shannon_entropy = pd.read_csv("data/pfeature/ha_avp_ic50_shannon_entrop
y.csv")
shannon_entropy = shannon_entropy.drop(['ID', 'Sequence'], axis=1)
residue_repeats = pd.read_csv("data/pfeature/ha_avp_ic50_residue_repeat.
csv")
residue_repeats = residue_repeats.drop(['ID', 'Sequence'], axis=1).reset
_index(drop=True)
sec_struct = pd.read_csv("data/pfeature/ha_avp_ic50_sec_struct.csv")
sec_struct = sec_struct.drop(['ID'], axis=1)
```

In [13]:
```
from Bio.SeqUtils.ProtParam import ProteinAnalysis
```

In [14]:
```
params = ['molecular_weight', 'aromaticity', 'instability_index',
          'isoelectric_point', 'helix', 'turn', 'sheet', 'with_reduced_c
ysteines',
          'with_disulfid_bridges', 'gravy', 'net_charge_at_pH7point4']
```

In [15]:
```
prop = []
for seq in df.Sequence:
    X = ProteinAnalysis(seq)
    molecular_weight = X.molecular_weight()
    aromaticity = X.aromaticity()
    instability_index = X.instability_index()
    isoelectric_point = X.isoelectric_point()
    sec_struc = X.secondary_structure_fraction()
    helix = sec_struc[0]
    turn = sec_struc[1]
    sheet = sec_struc[2]
    epsilon_prot = X.molar_extinction_coefficient()
    with_reduced_cysteines = epsilon_prot[0]
    with_disulfid_bridges = epsilon_prot[1]
    gravy = X.gravy() # hydrophobicity related
    # flexibility = X.flexibility()
    # X.protein_scale()
    net_charge_at_pH7point4 = X.charge_at_pH(7.4)

    prop.append([molecular_weight, aromaticity, instability_index, isoel
ectric_point, helix, turn, sheet,
                 with_reduced_cysteines, with_disulfid_bridges, gravy, n
et_charge_at_pH7point4])
prop = pd.DataFrame(prop, columns=params)
```

In [17]:
```
physicochem_prop = pd.concat([prop, shannon_entropy, residue_repeats, se
c_struct], axis=1)
```

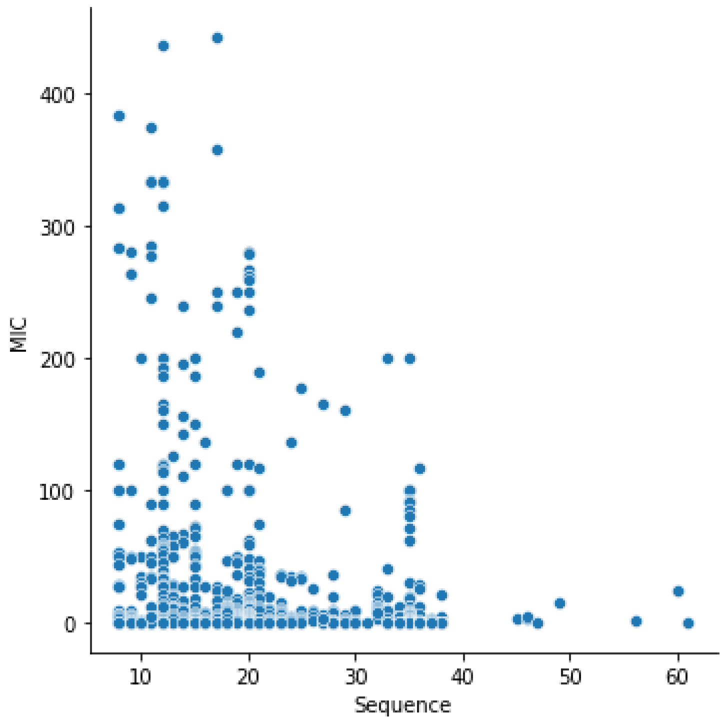
In [50]:
```
seq_len = df['Sequence'].map(lambda s: len(s))
```

In [51]:
```
np.corrcoef(seq_len, df['MIC'])
```

Out[51]:
```
array([[ 1.        , -0.27406559],
       [-0.27406559,  1.        ]])
```

In [53]:
```
sns.relplot(x=seq_len, y=df['MIC'])
```

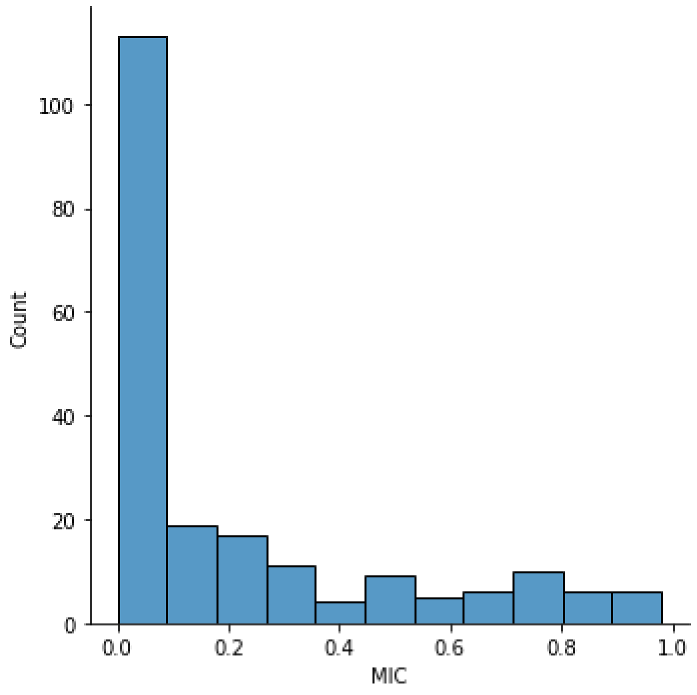
Out[53]:
```
<seaborn.axisgrid.FacetGrid at 0x1a22cd6438>
```



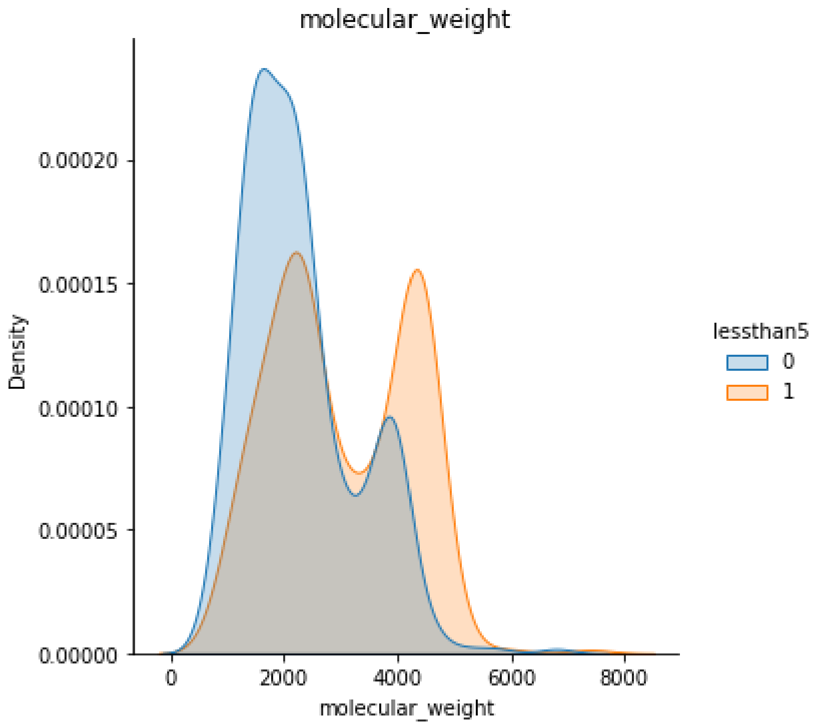
In [58]:
```
df2 = df[df['MIC']<1]
```

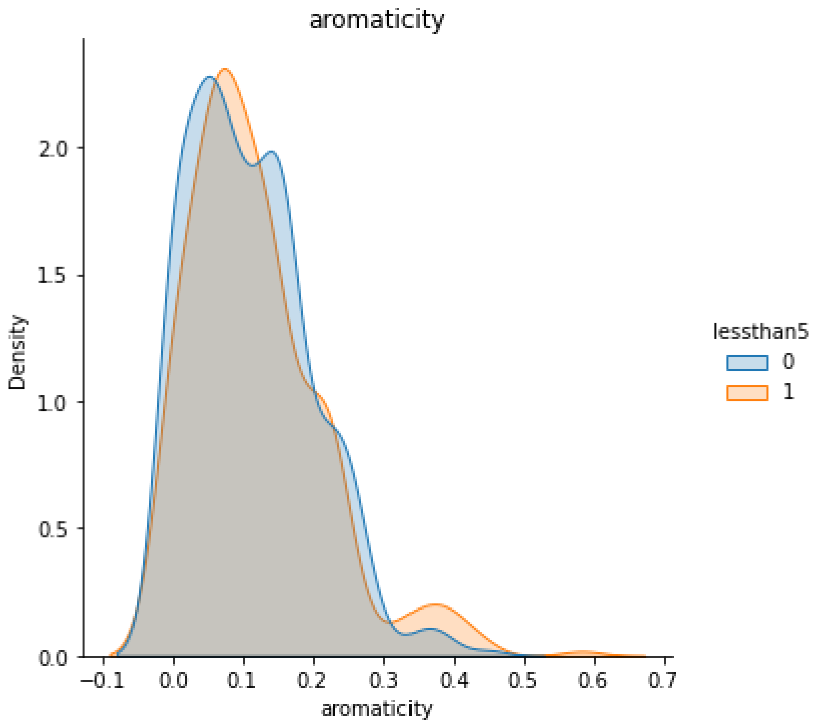
In [59]:
```
sns.displot(x=df2['MIC'])
```

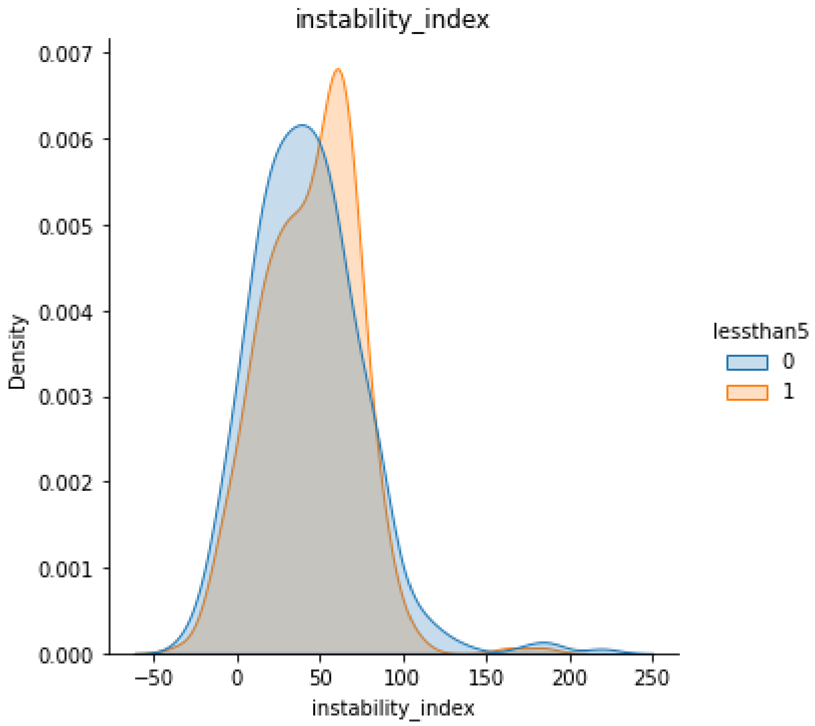
Out[59]:
```
<seaborn.axisgrid.FacetGrid at 0x1a217d0be0>
```

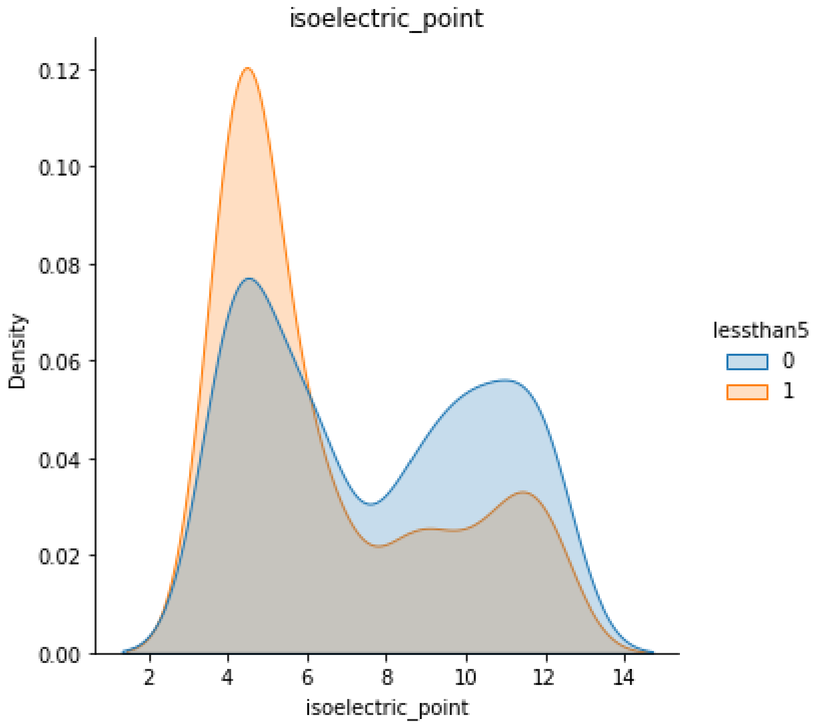


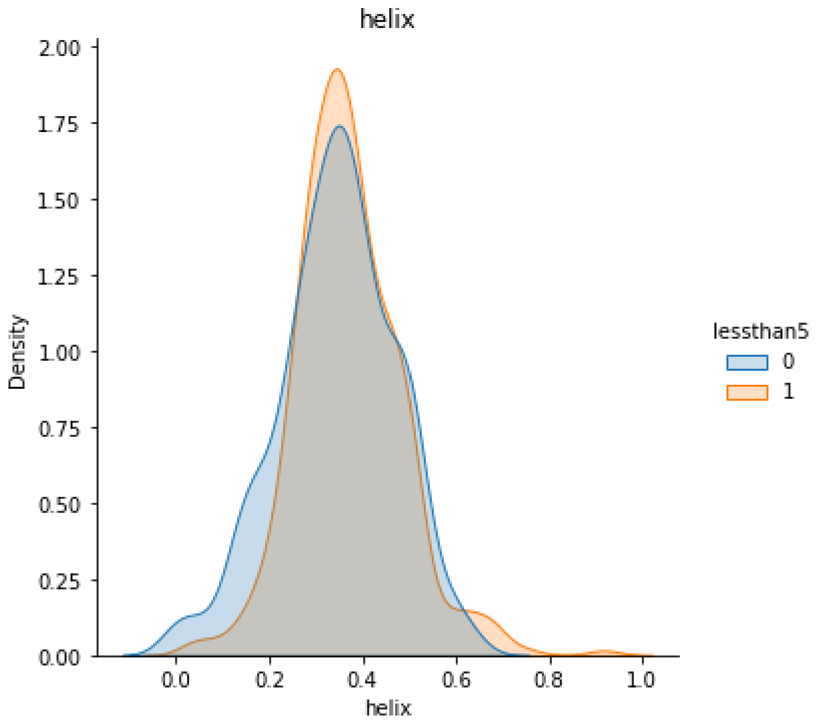
In [23]:
```
for feature in physicochem_prop.columns:
    sns.displot(x=physicochem_prop[feature], hue=df['lessthan5'], kind=
"kde", fill=True)
    plt.title(feature)
    plt.show()
```

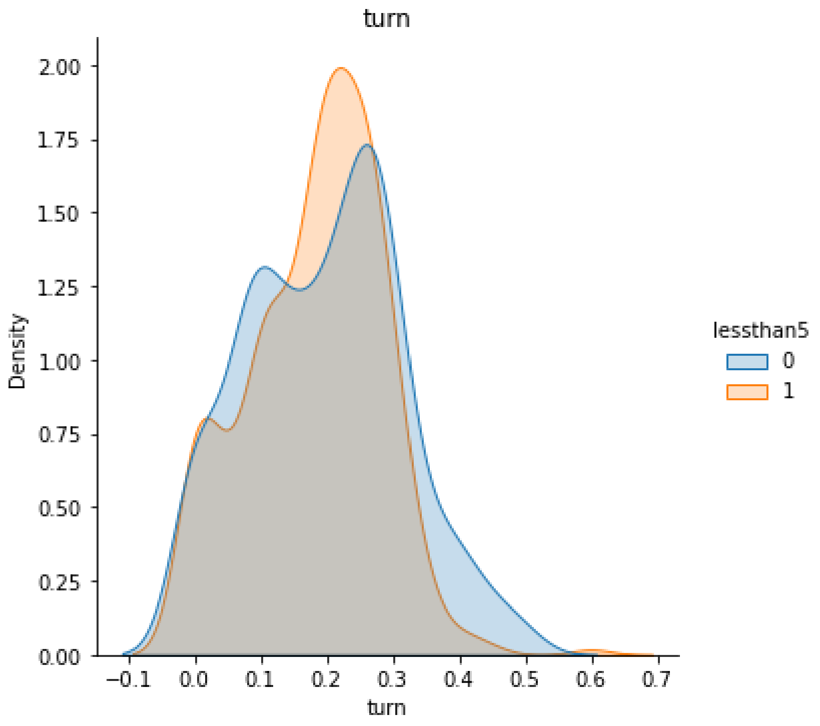

molecular_weight
Density
0.00000
0.00005
0.00010
0.00015
0.00020
0
2000
4000
6000
8000
molecular_weight
lessthan5
0
1

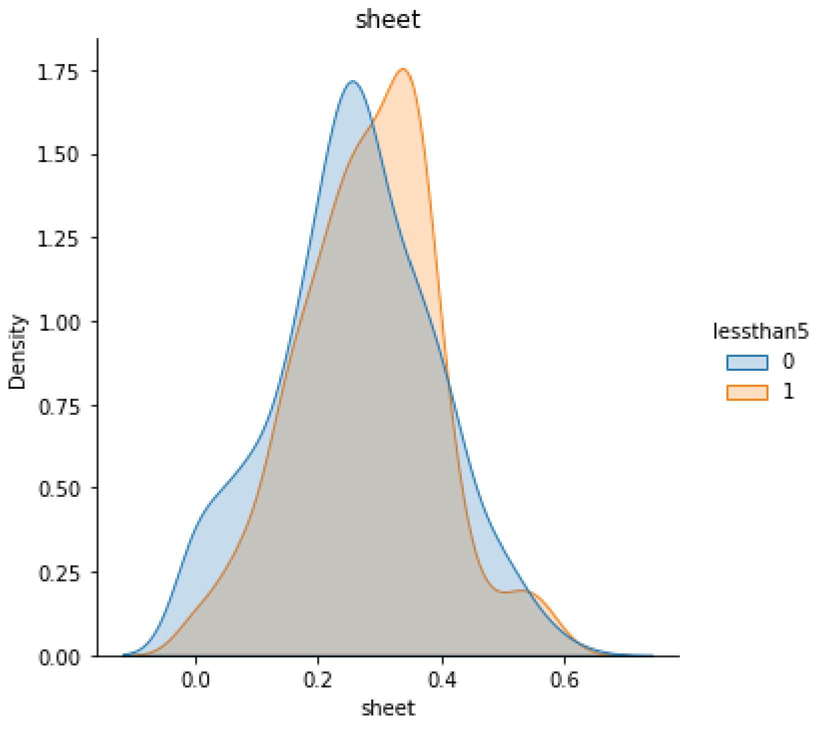

aromaticity
Density
0.0
0.5
1.0
1.5
2.0
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
aromaticity
lessthan5
0
1

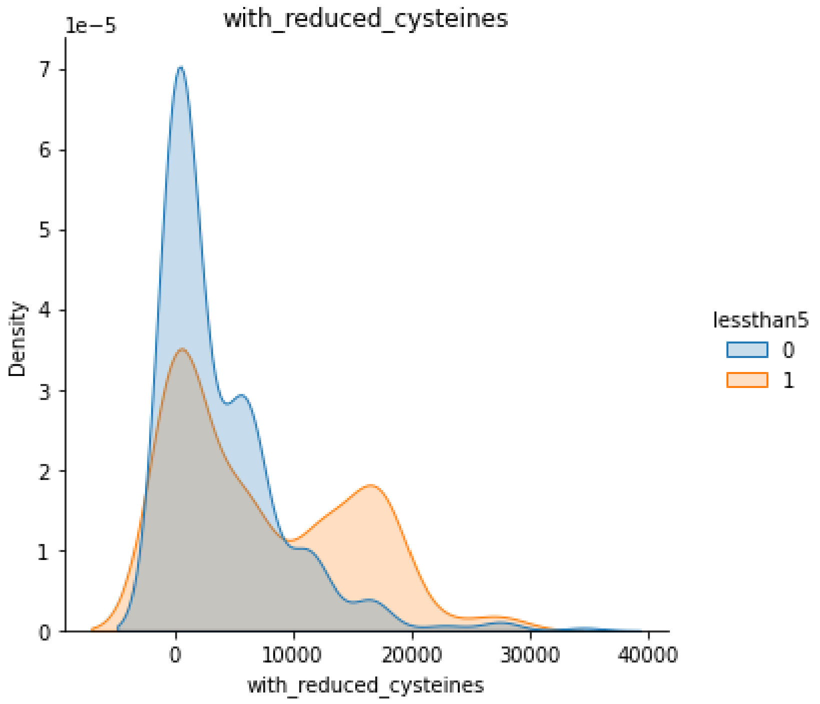

instability_index
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
−50
0
50
100
150
200
250
instability_index
lessthan5
0
1

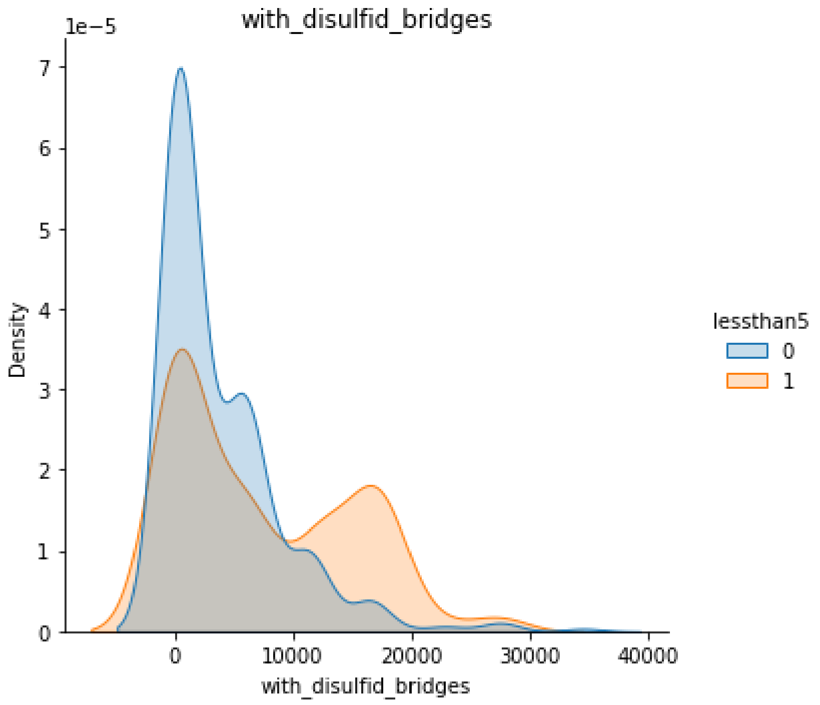

isoelectric_point
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
2
4
6
8
10
12
14
isoelectric_point
lessthan5
0
1

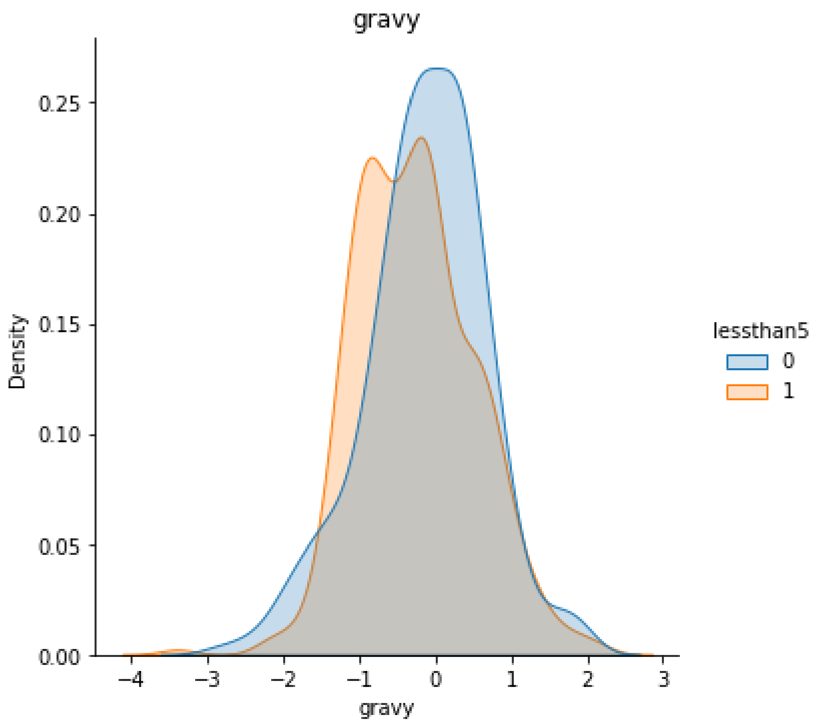

helix
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
helix
lessthan5
0
1

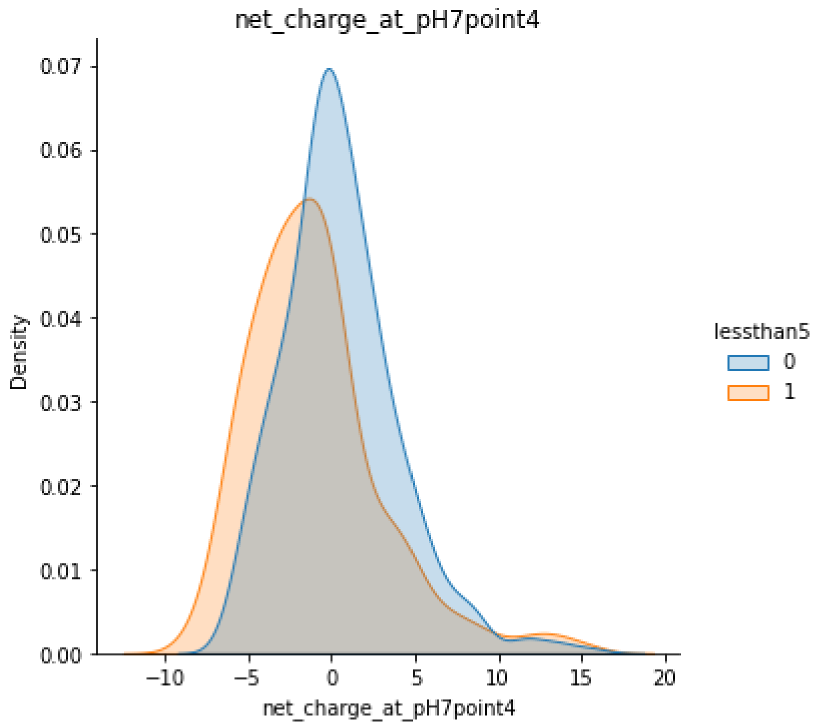

turn
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
turn
lessthan5
0
1

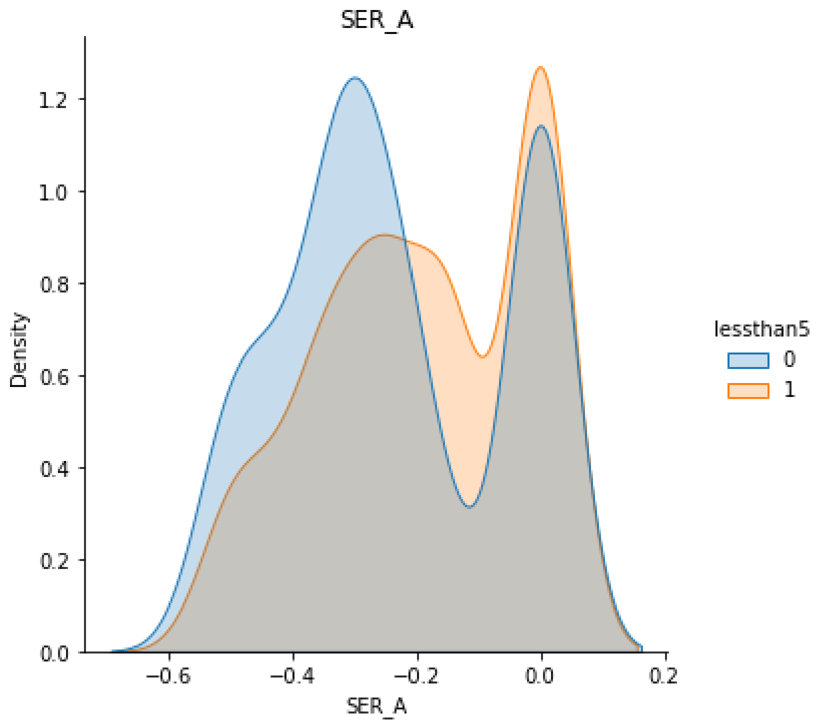

sheet
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
sheet
lessthan5
0
1

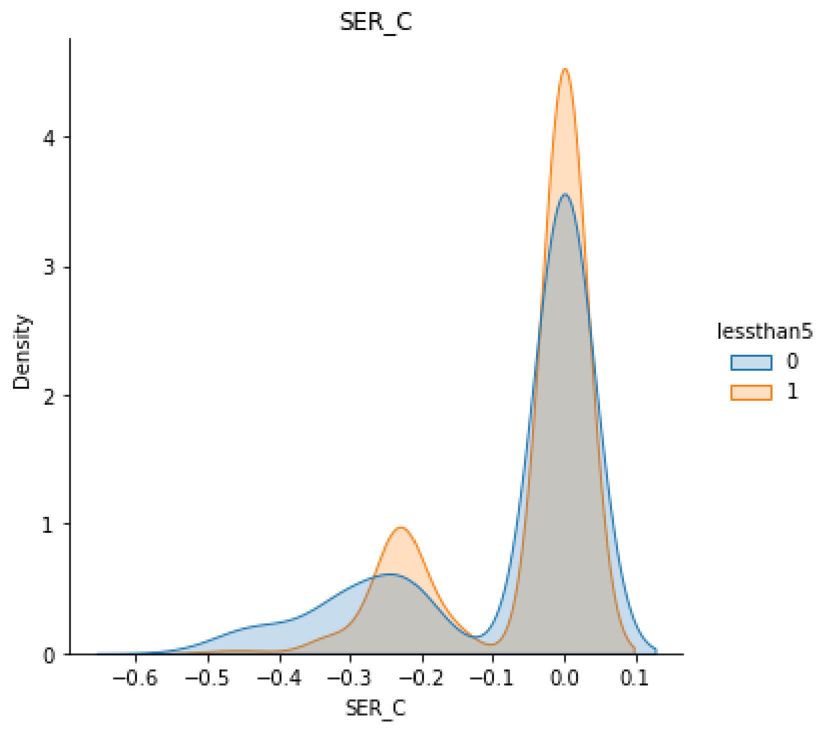

with_reduced_cysteines
1e−5
Density
7
6
5
4
3
2
1
0
0
10000
20000
30000
40000
with_reduced_cysteines
lessthan5
0
1

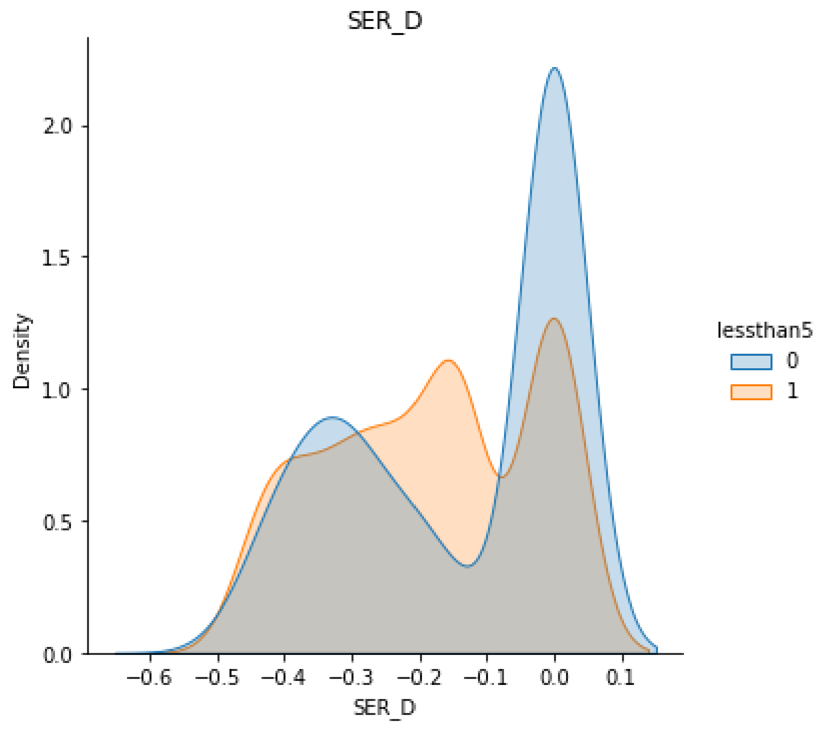

with_disulfid_bridges
1e−5
7
6
5
4
3
2
1
0
Density
0
10000
20000
30000
40000
with_disulfid_bridges
lessthan5
0
1

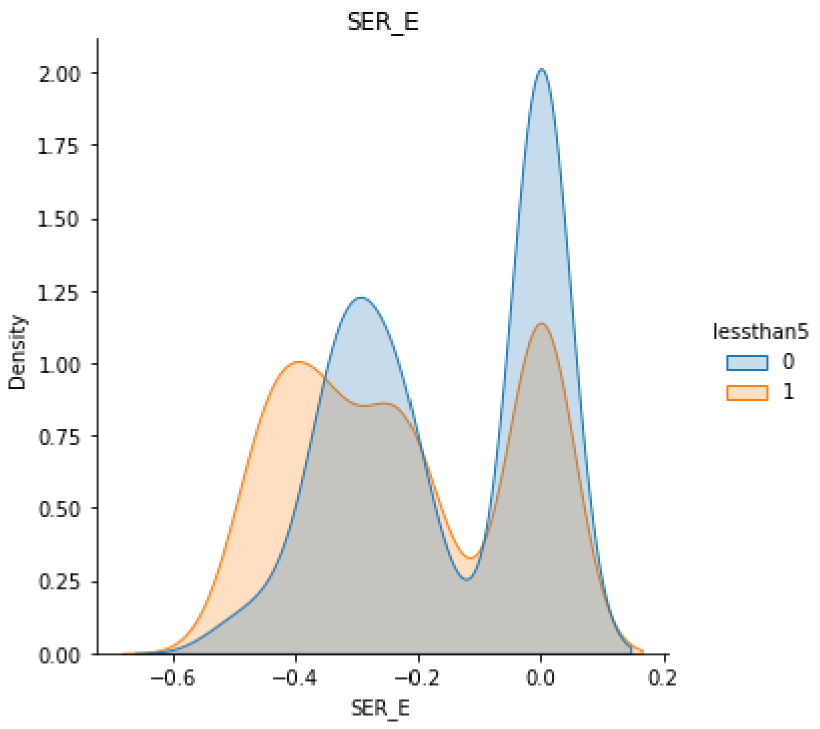

gravy
0.25
0.20
0.15
0.10
0.05
0.00
Density
−4
−3
−2
−1
0
1
2
3
gravy
lessthan5
0
1

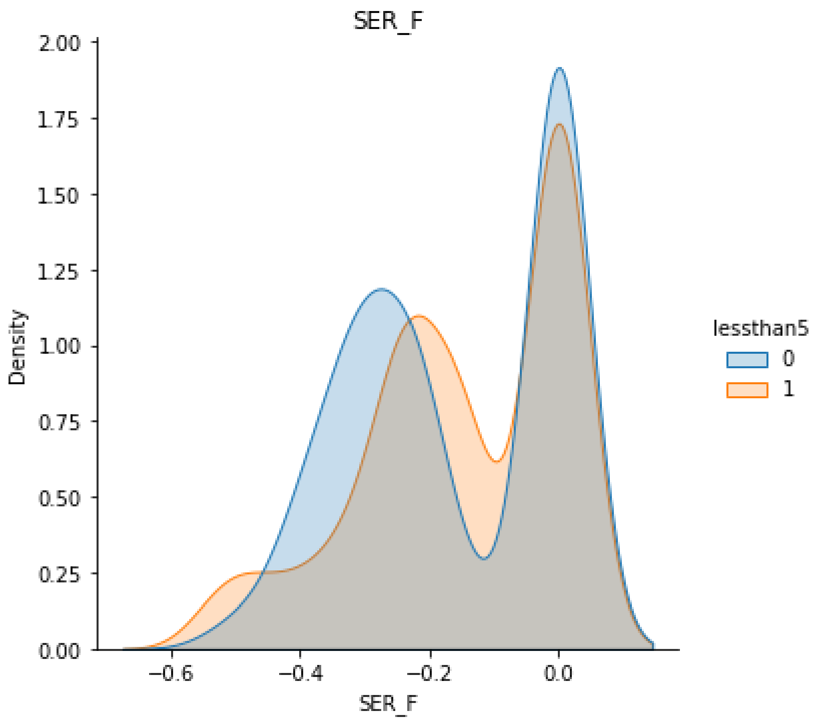

net_charge_at_pH7point4
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−10
−5
0
5
10
15
20
net_charge_at_pH7point4
lessthan5
0
1

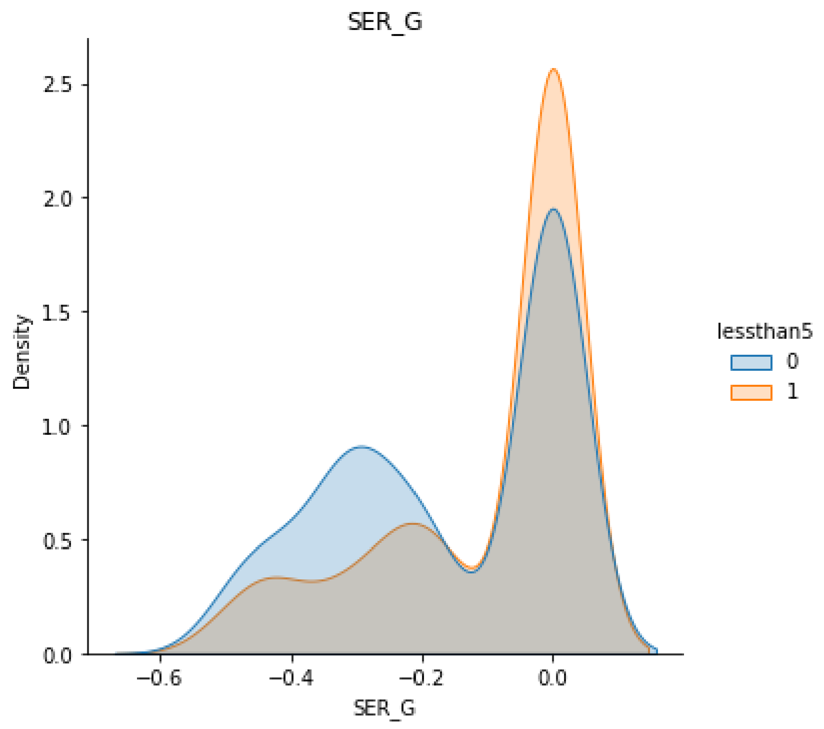

SER_A
Density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.6
−0.4
−0.2
0.0
0.2
SER_A
lessthan5
0
1

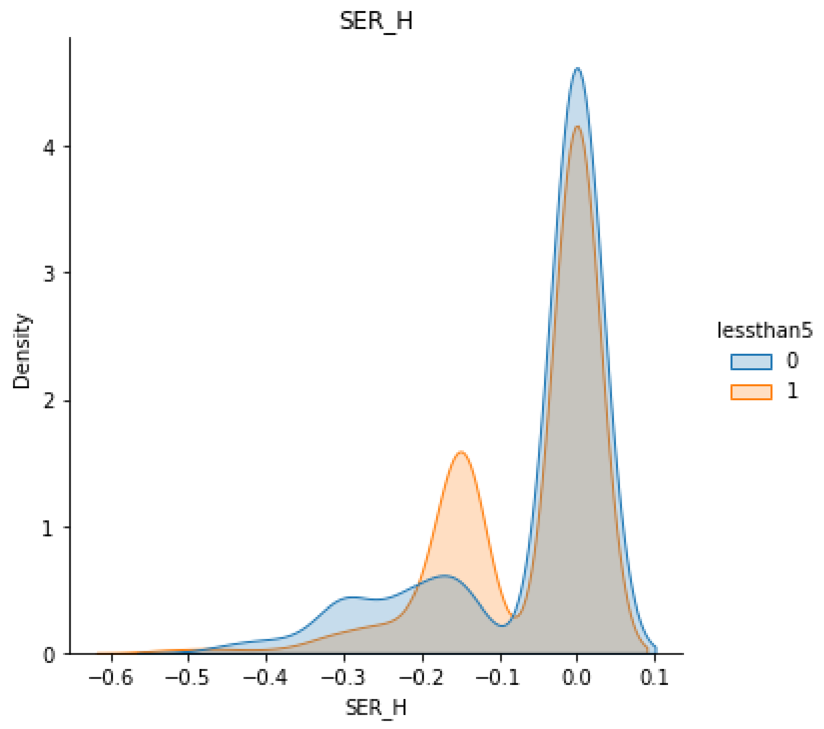

SER_C
Density
SER_C
lessthan5
0
1
4
3
2
1
0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1

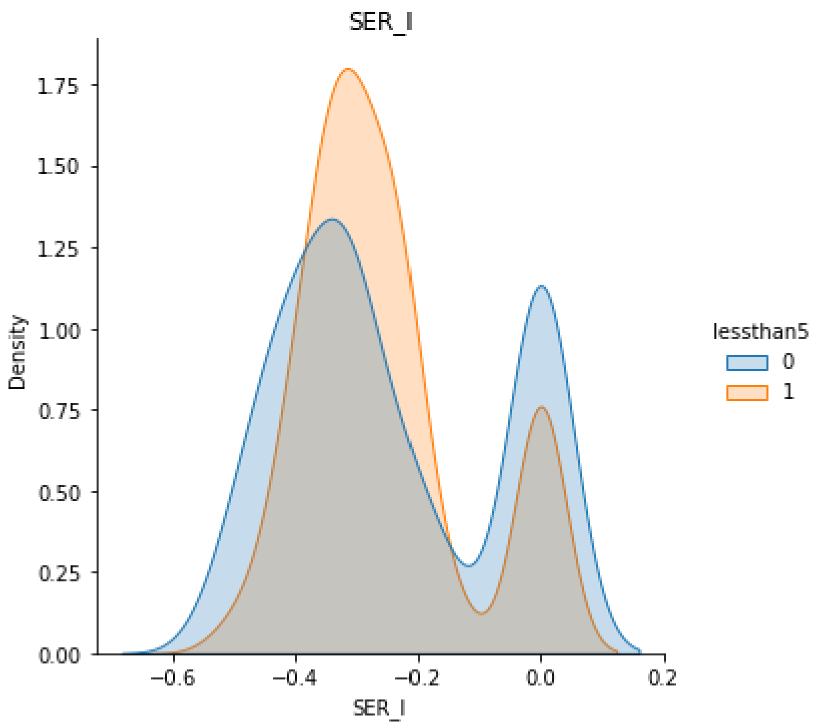

SER_D
Density
SER_D
lessthan5
0
1
2.0
1.5
1.0
0.5
0.0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1

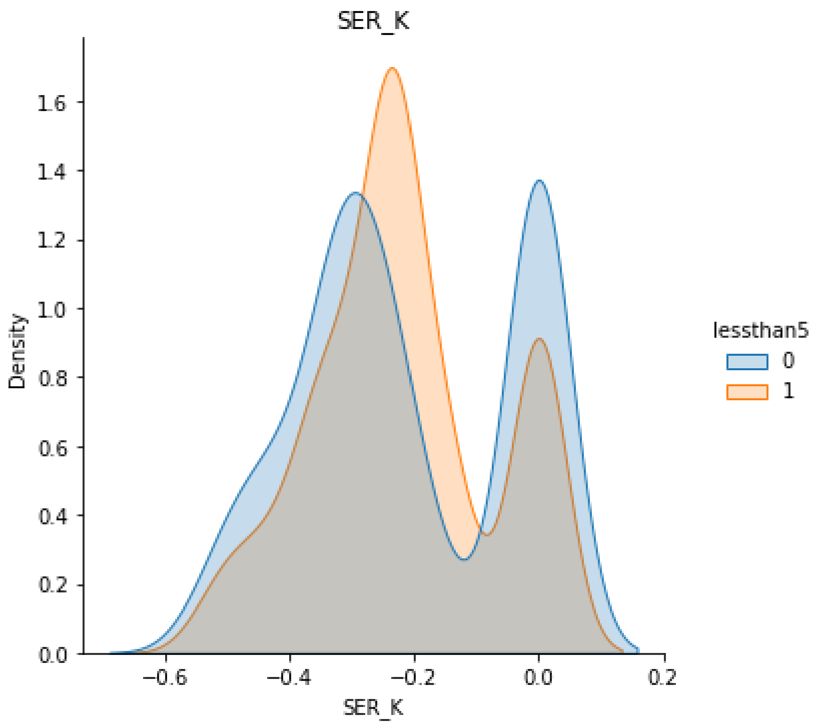

SER_E
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.6
−0.4
−0.2
0.0
0.2
SER_E
lessthan5
0
1

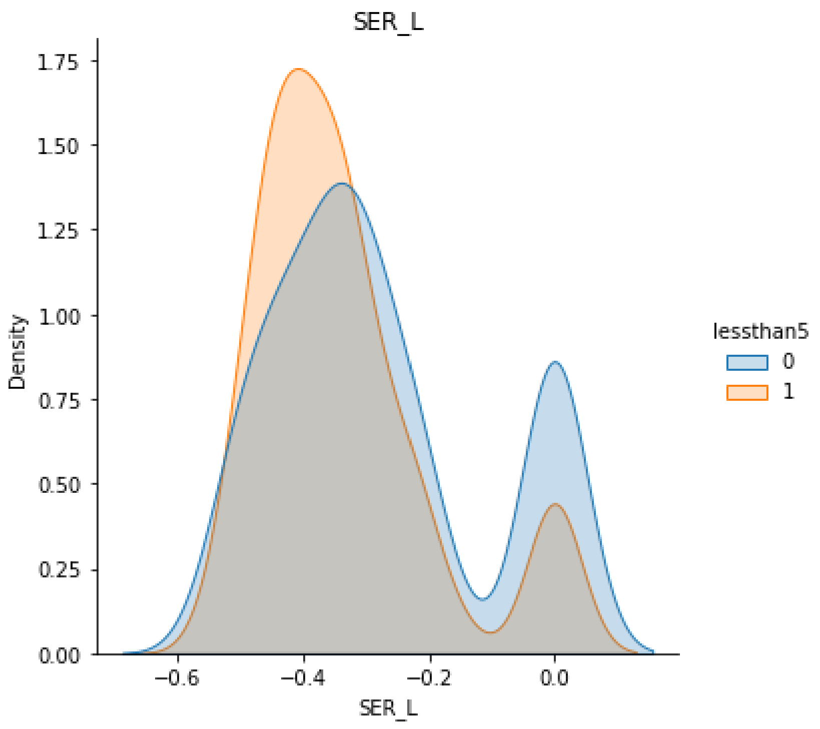

SER_F
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.6
−0.4
−0.2
0.0
SER_F
lessthan5
0
1

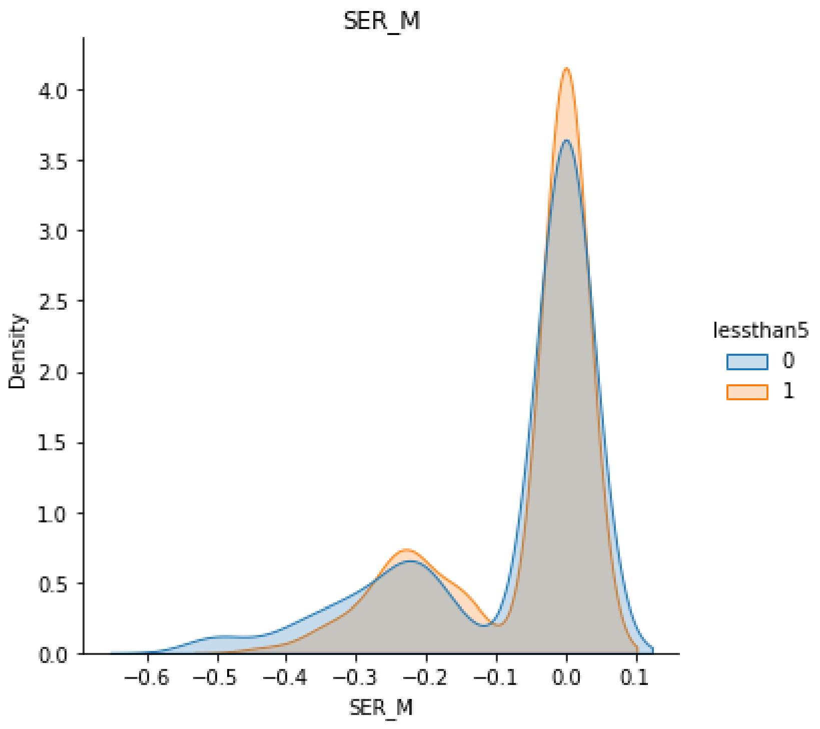

SER_G
Density
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0
SER_G
lessthan5
0
1

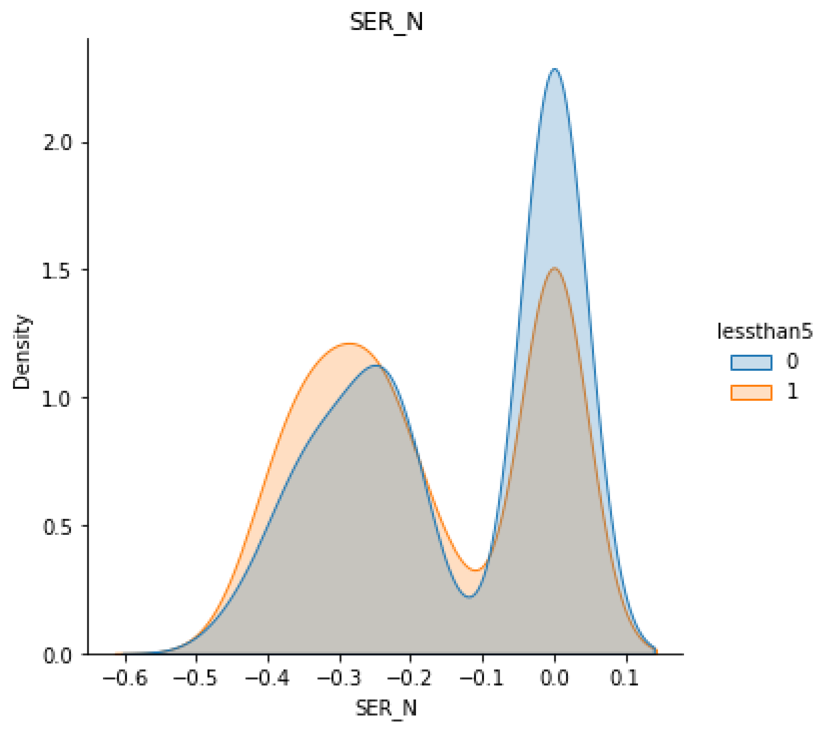

SER_H
Density
4
3
2
1
0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
SER_H
lessthan5
0
1

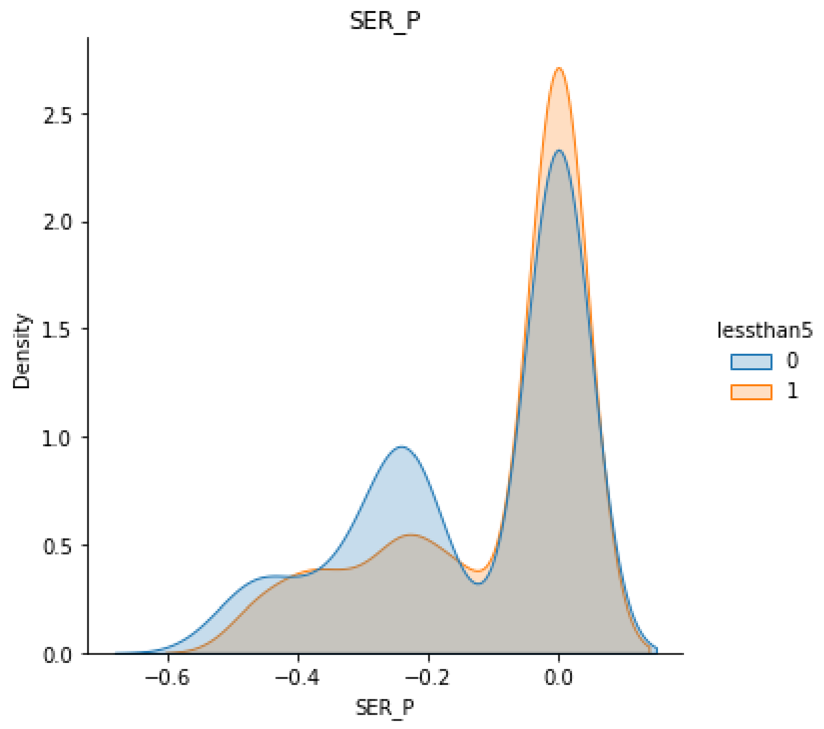

SER_I
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.6
−0.4
−0.2
0.0
0.2
SER_I
lessthan5
0
1

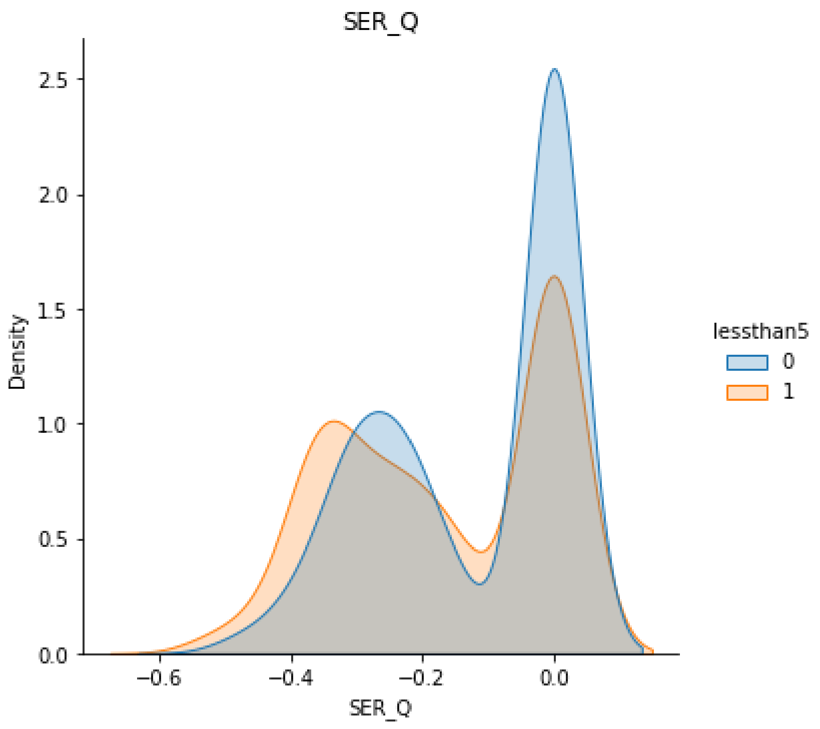

SER_K
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.6
−0.4
−0.2
0.0
0.2
SER_K
lessthan5
0
1

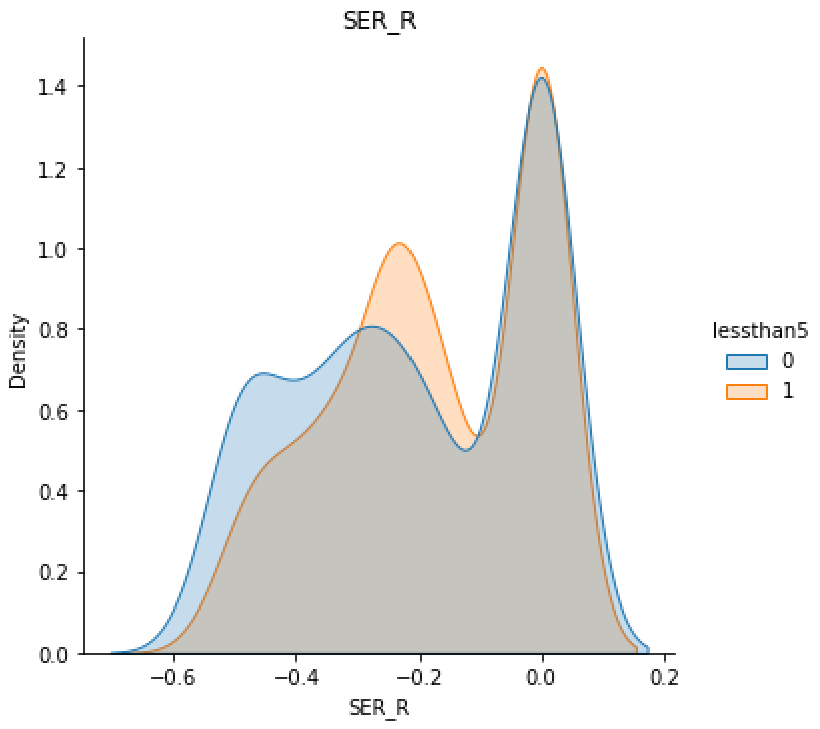

SER_L
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.6
−0.4
−0.2
0.0
SER_L
lessthan5
0
1

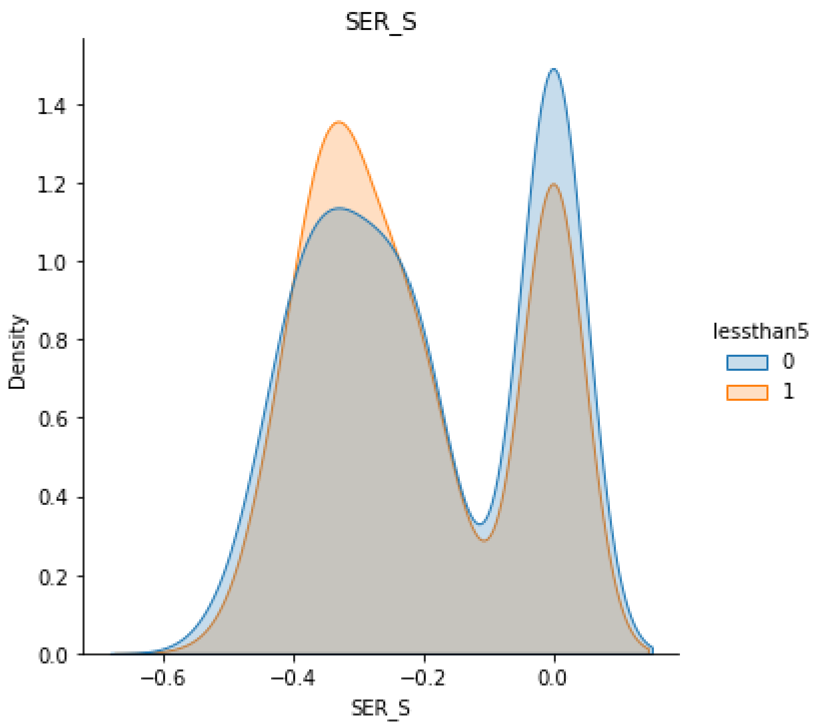

SER_M
Density
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
SER_M
lessthan5
0
1

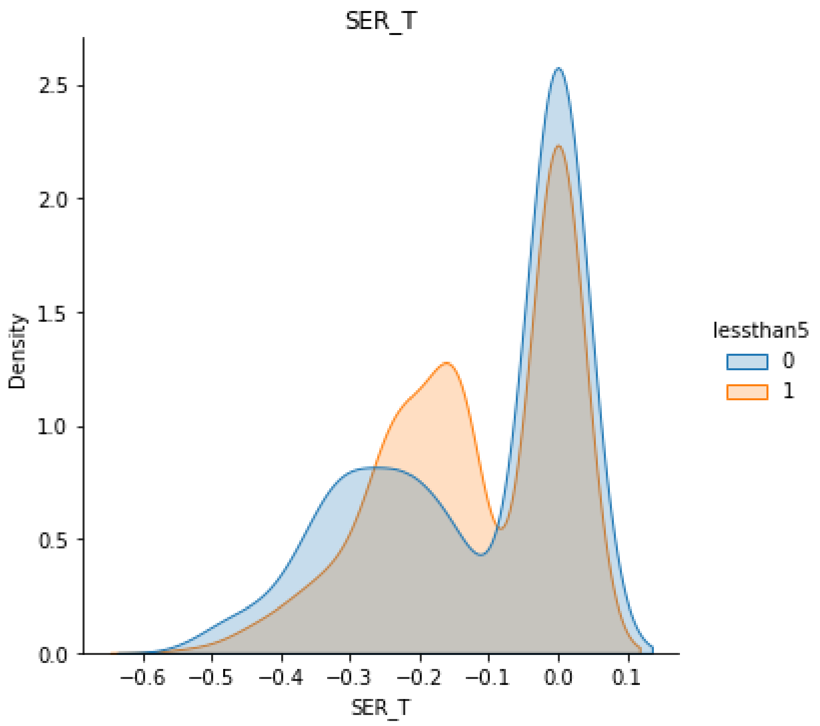

SER_N
Density
SER_N
lessthan5
0
1
0.0
0.5
1.0
1.5
2.0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1

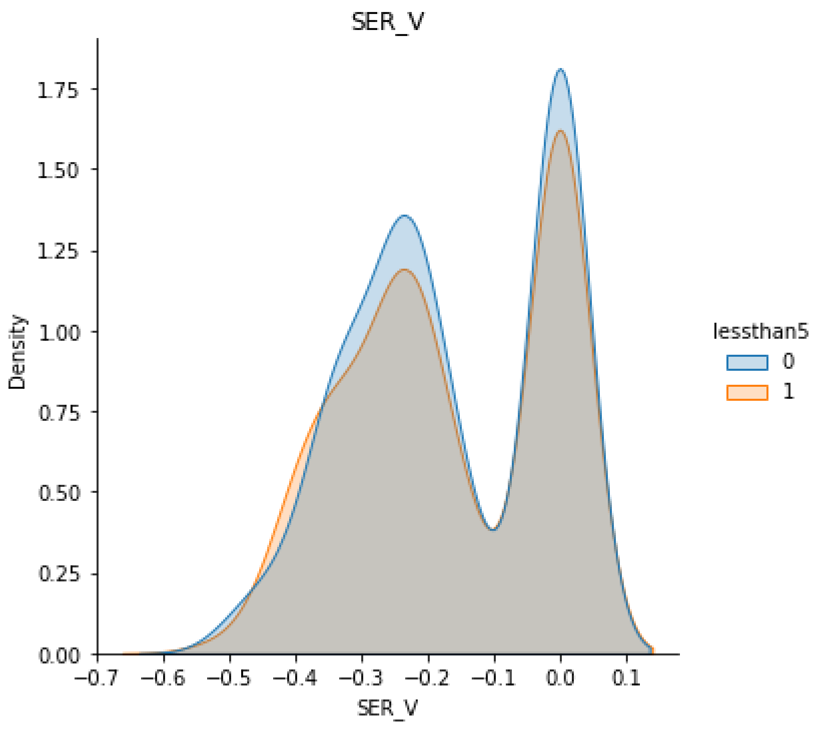

SER_P
Density
SER_P
lessthan5
0
1
0.0
0.5
1.0
1.5
2.0
2.5
−0.6
−0.4
−0.2
0.0

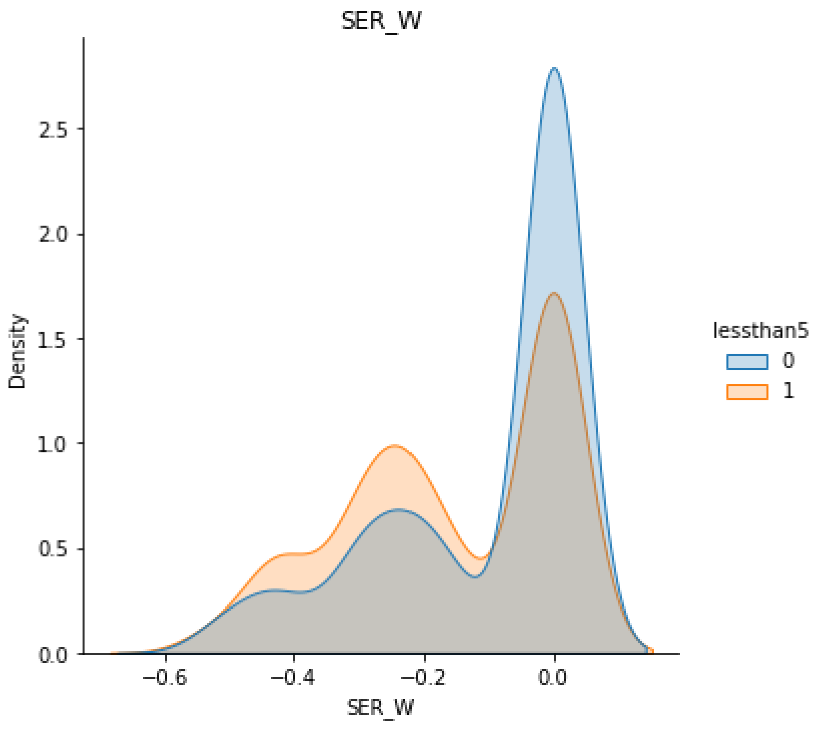

SER_Q
Density
SER_Q
lessthan5
0
1
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0

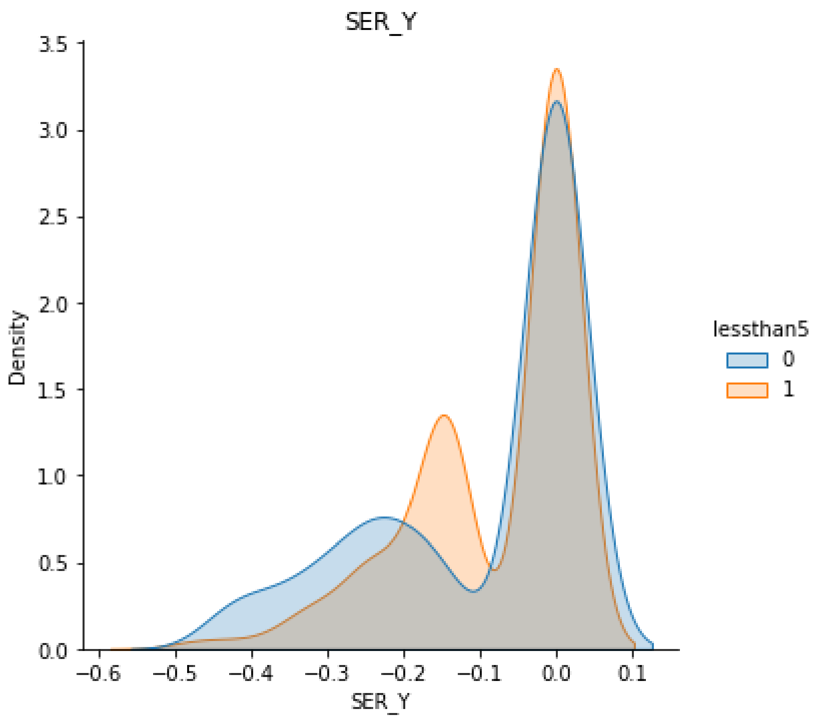

SER_R
Density
SER_R
lessthan5
0
1
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.6
−0.4
−0.2
0.0
0.2

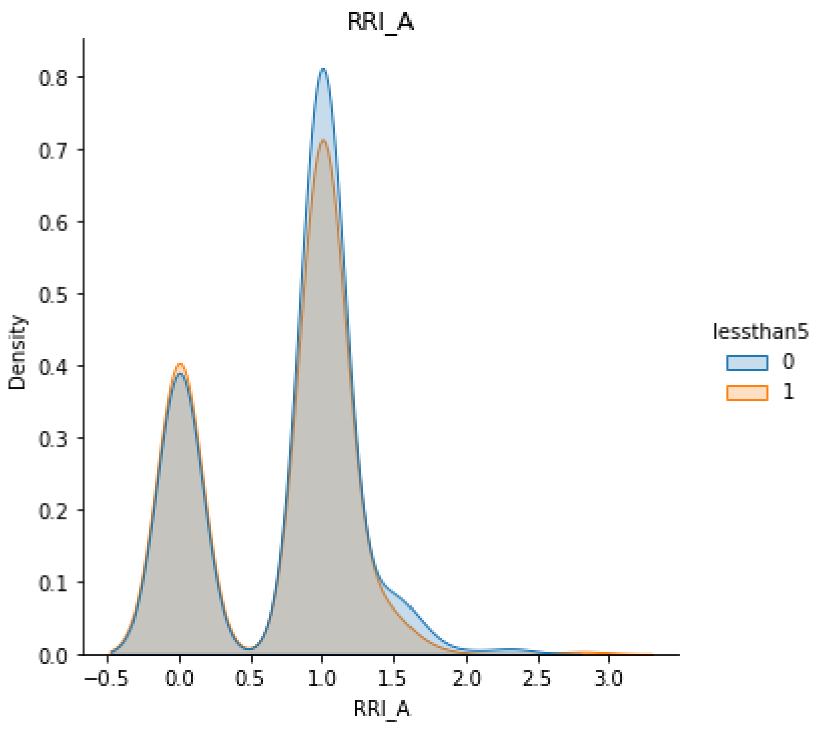

SER_S
Density
SER_S
lessthan5
0
1
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.6
−0.4
−0.2
0.0

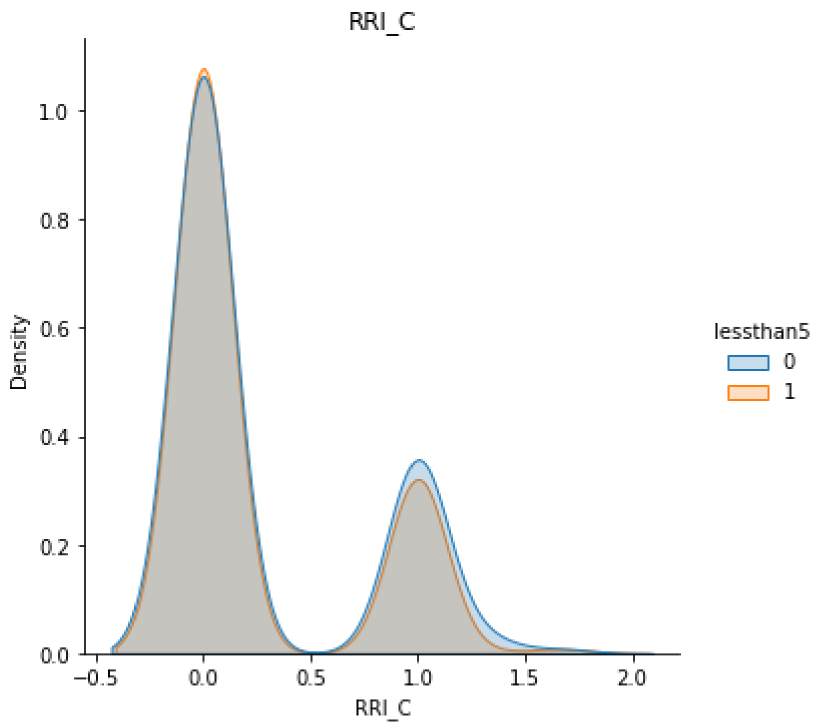

SER_T
Density
SER_T
lessthan5
0
1
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1


SER_V
Density
SER_V
lessthan5
0
1
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1

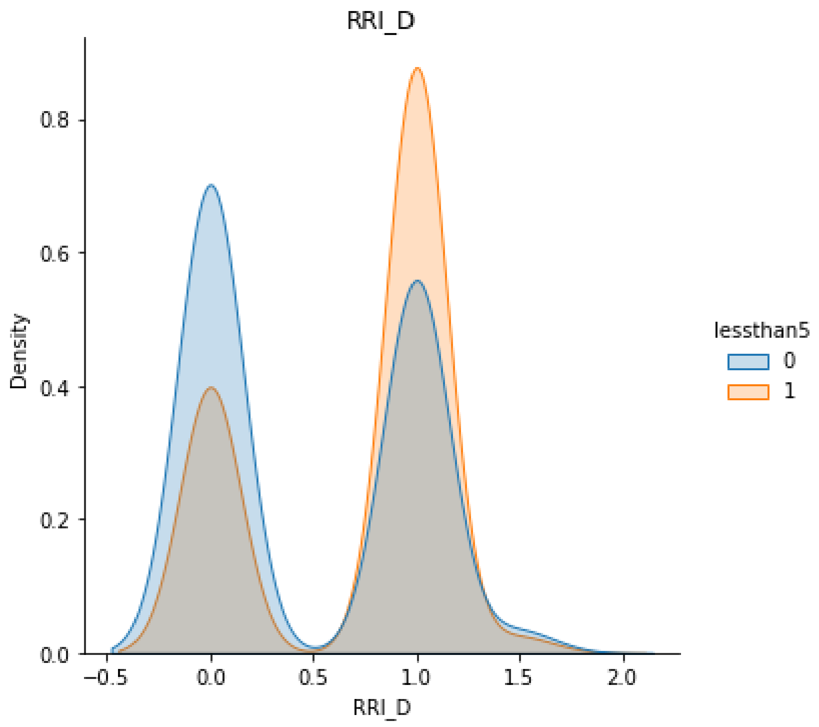

SER_W
Density
SER_W
lessthan5
0
1
2.5
2.0
1.5
1.0
0.5
0.0
−0.6
−0.4
−0.2
0.0


SER_Y
Density
SER_Y
lessthan5
0
1

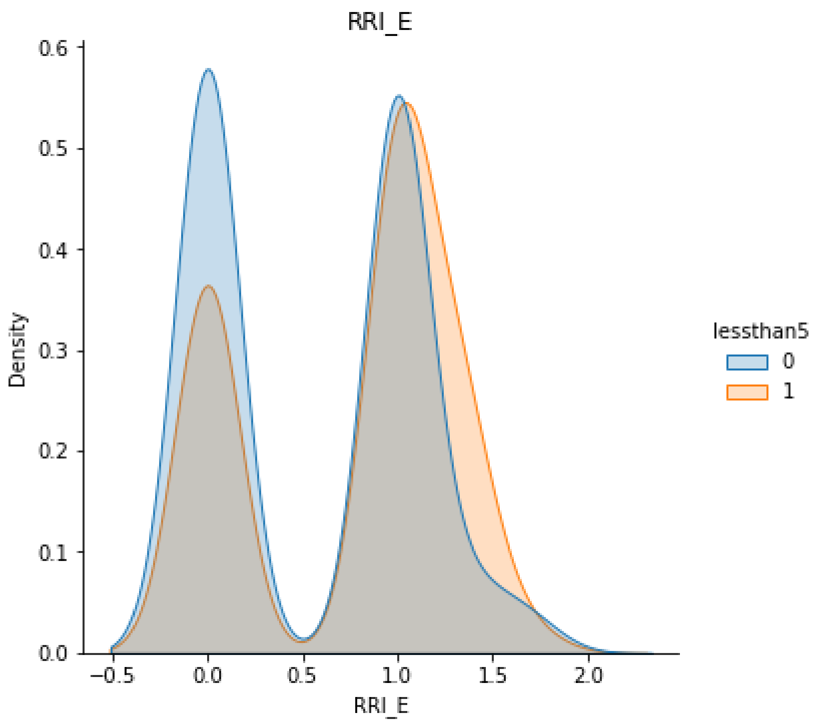

RRI_A
Density
RRI_A
lessthan5
0
1

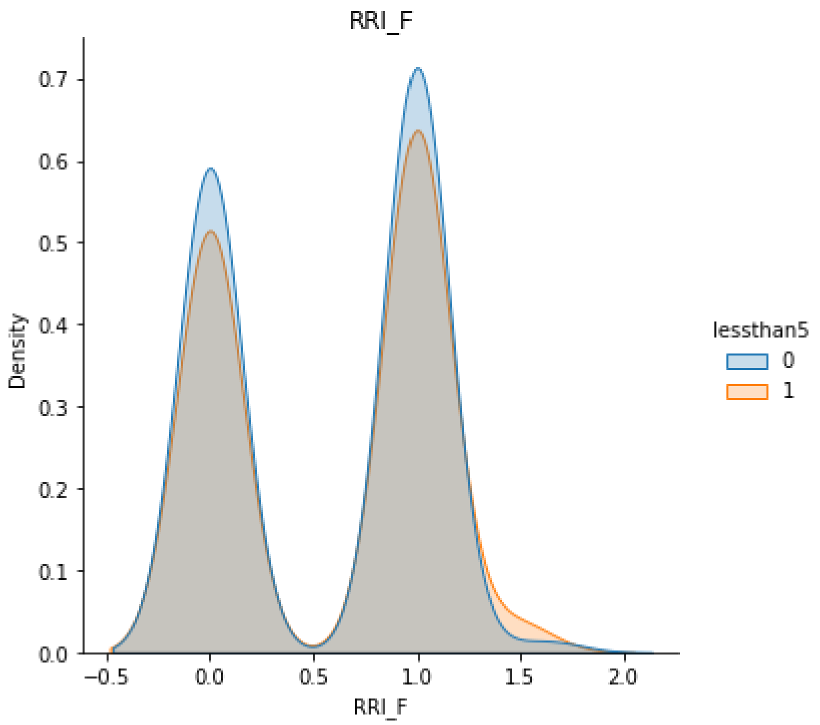


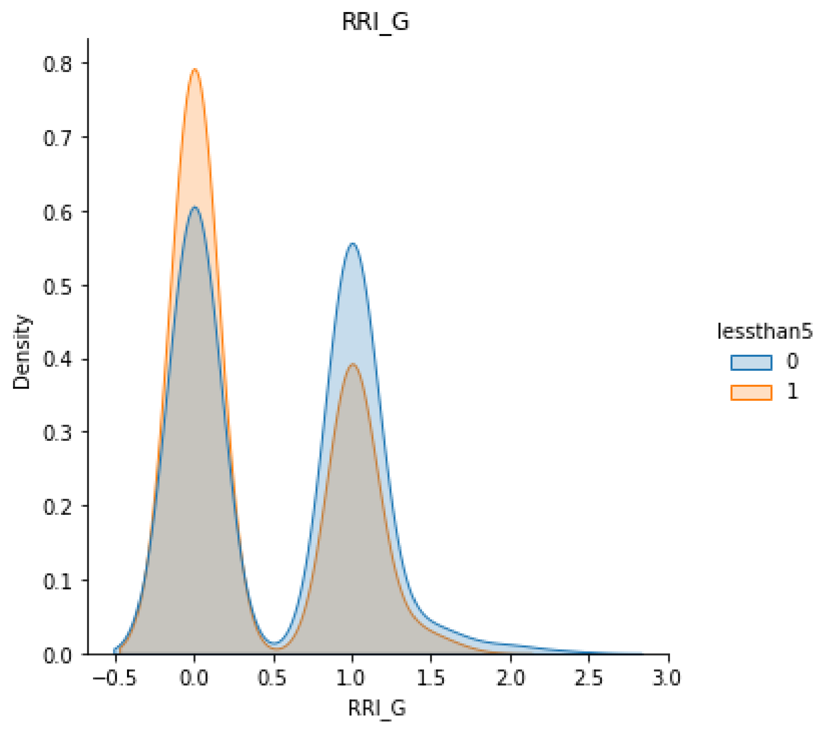
RRI_C
Density
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
0.0
0.5
1.0
1.5
2.0
RRI_C
lessthan5
0
1

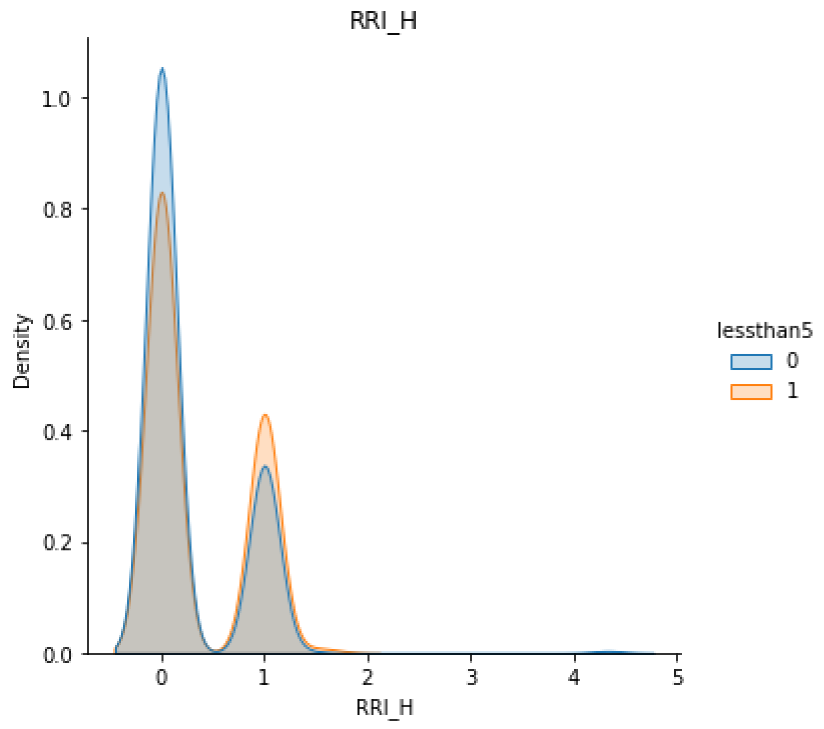


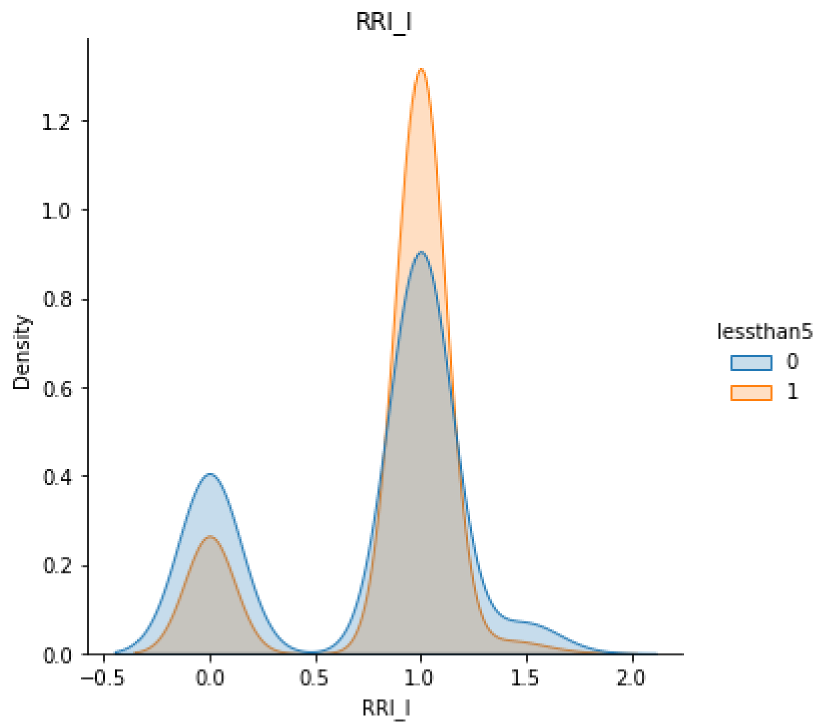
RRI_D
Density
0.8
0.6
0.4
0.2
0.0
−0.5
0.0
0.5
1.0
1.5
2.0
RRI_D
lessthan5
0
1

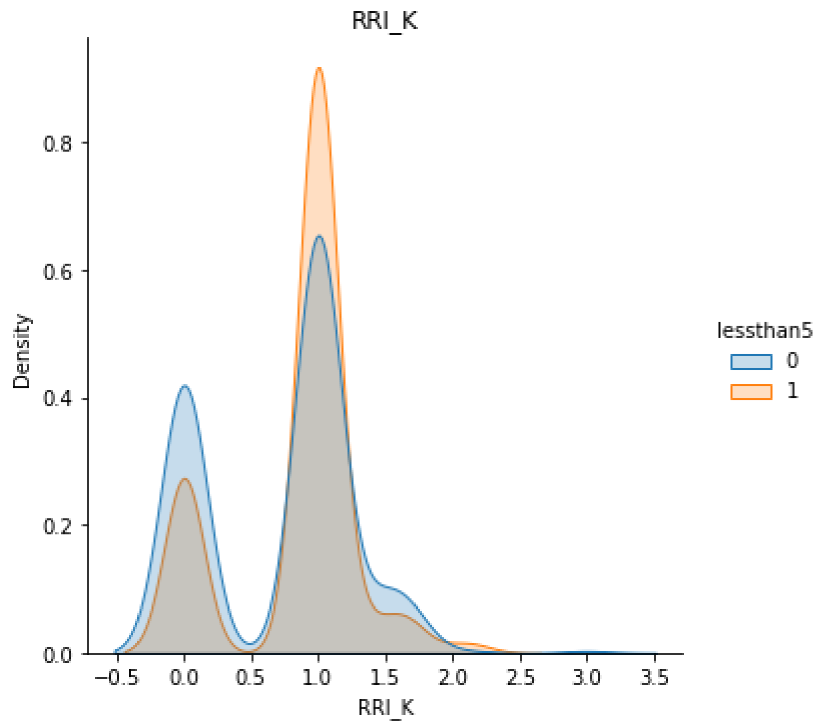

RRI_E
Density
RRI_E
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−0.5
0.0
0.5
1.0
1.5
2.0

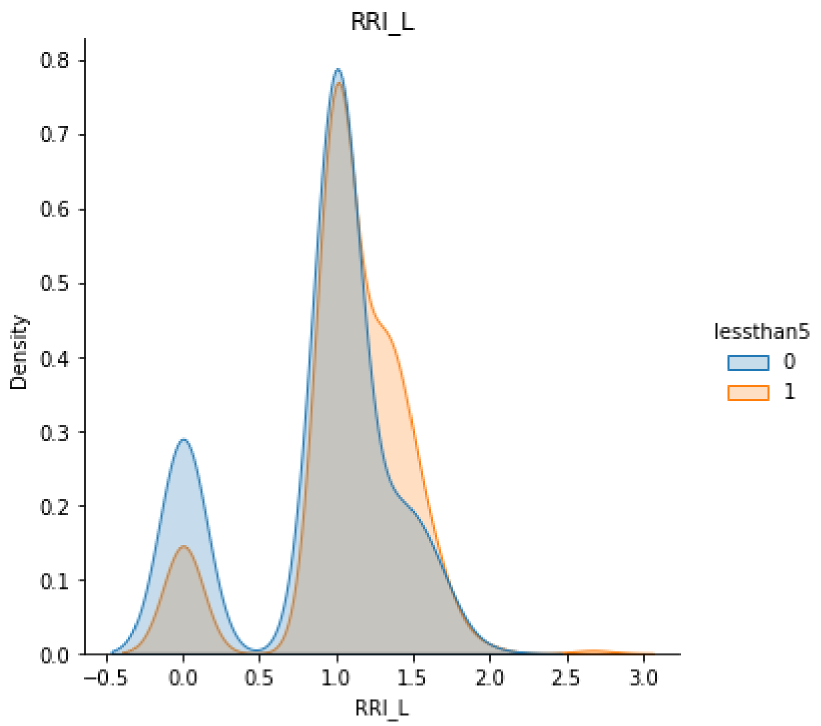

RRI_F
Density
RRI_F
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−0.5
0.0
0.5
1.0
1.5
2.0

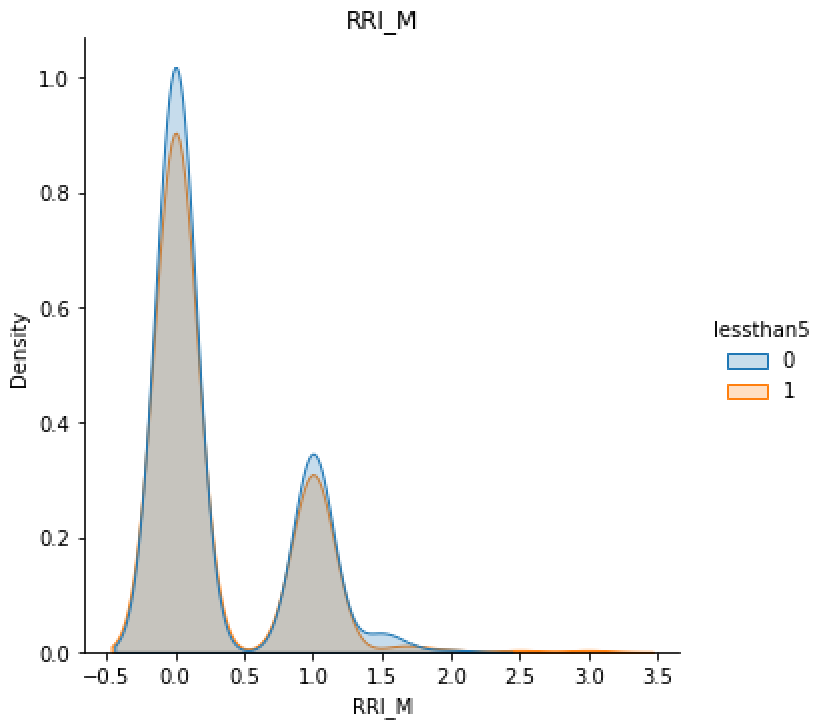

RRI_G
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RRI_G
lessthan5
0
1

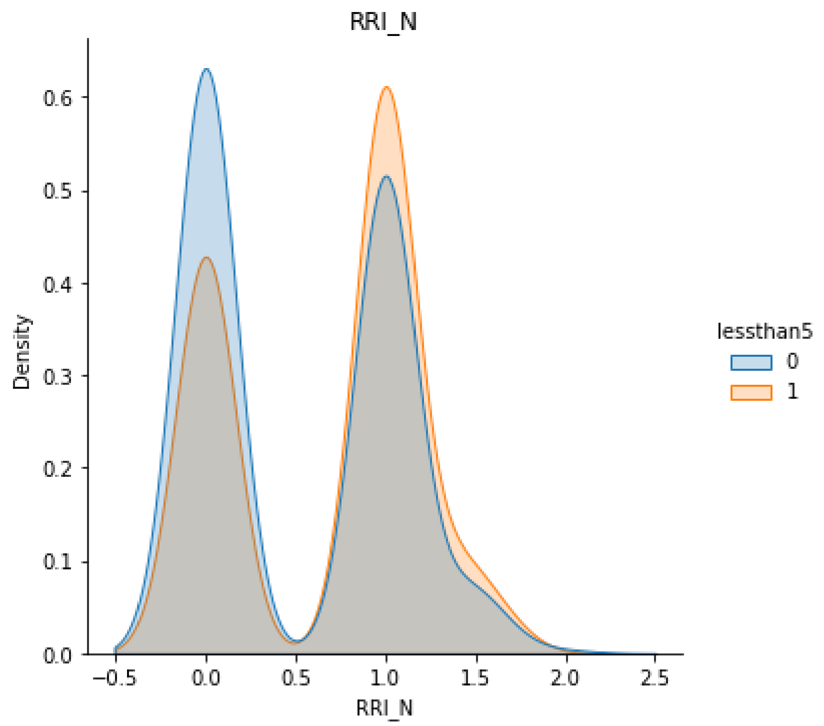

RRI_H
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
RRI_H
lessthan5
0
1

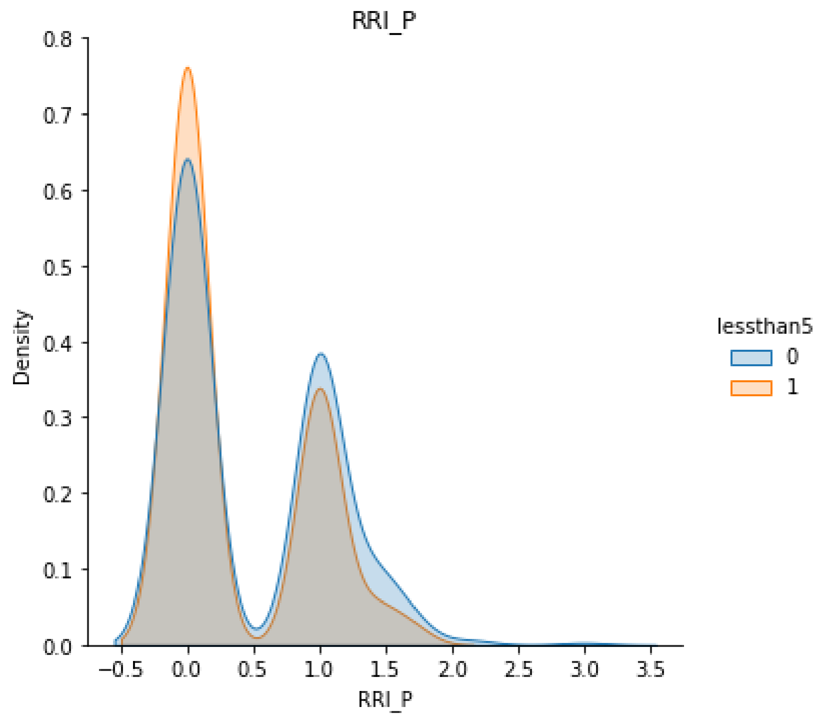

RRI_I
Density
RRI_I
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−0.5
0.0
0.5
1.0
1.5
2.0

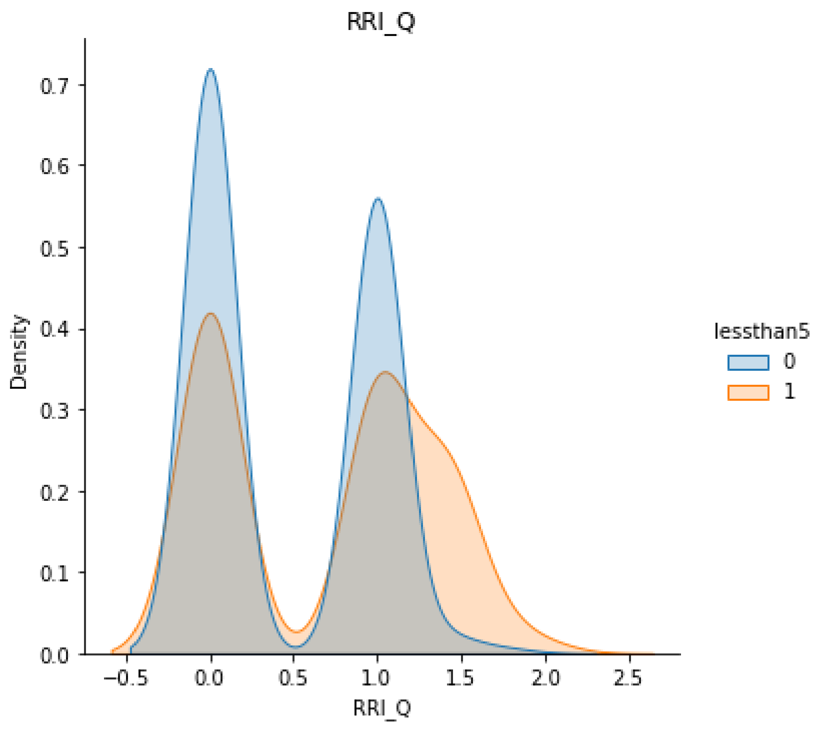

RRI_K
Density
RRI_K
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

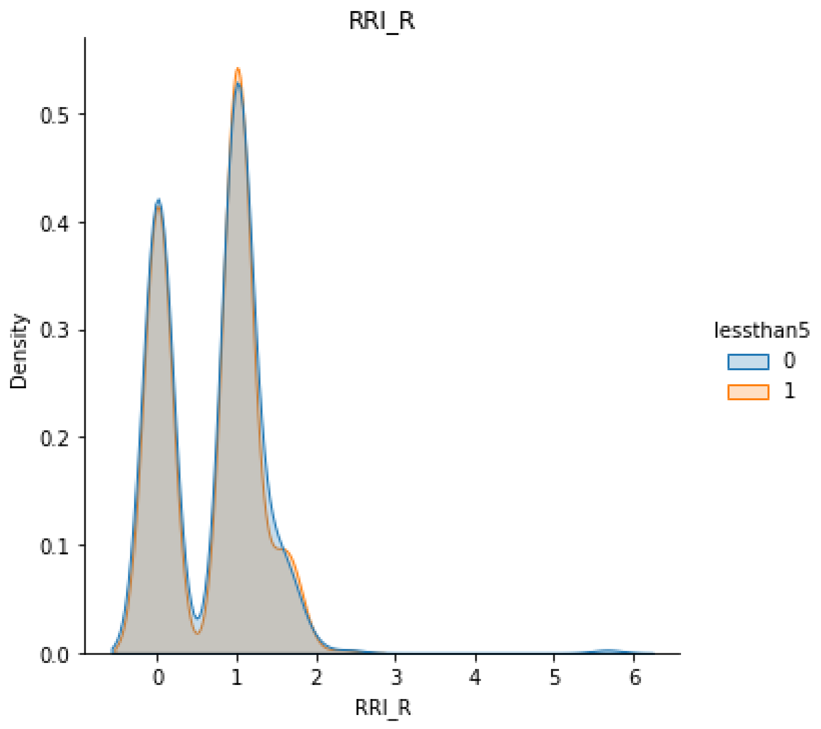

RRI_L
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RRI_L
lessthan5
0
1

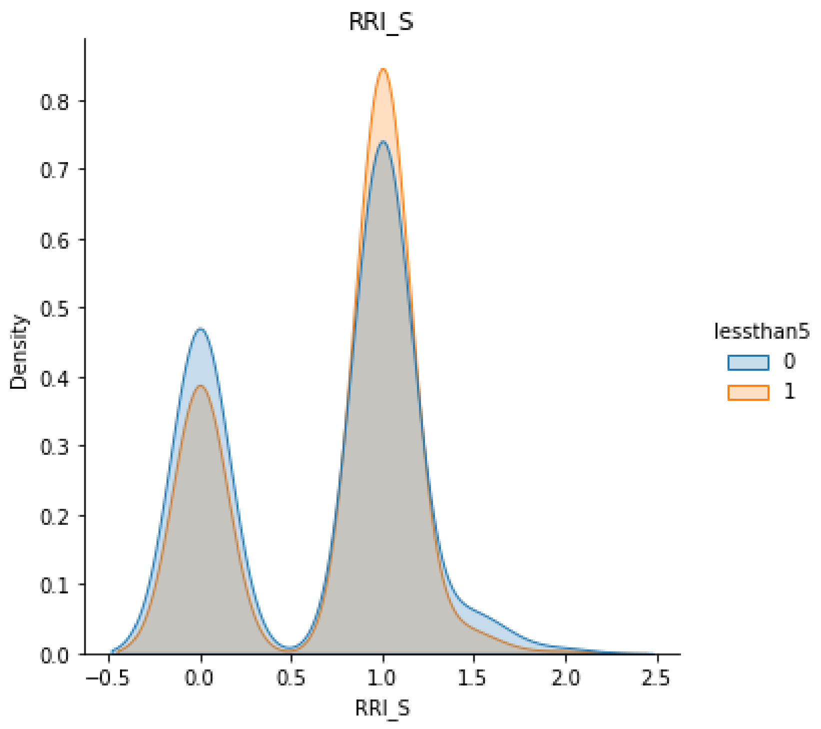

RRI_M
Density
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
RRI_M
lessthan5
0
1

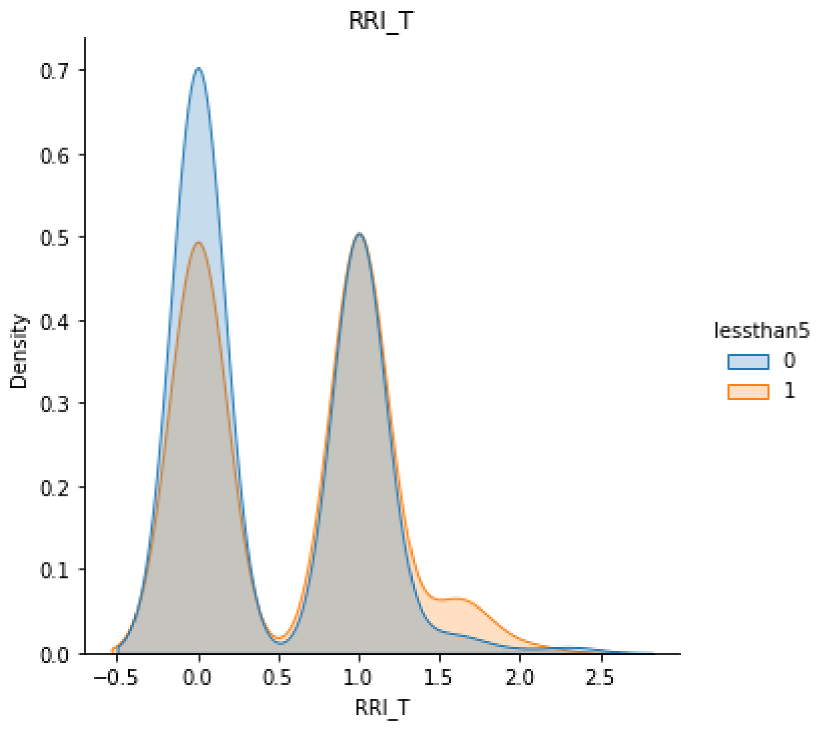

RRI_N
Density
RRI_N
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

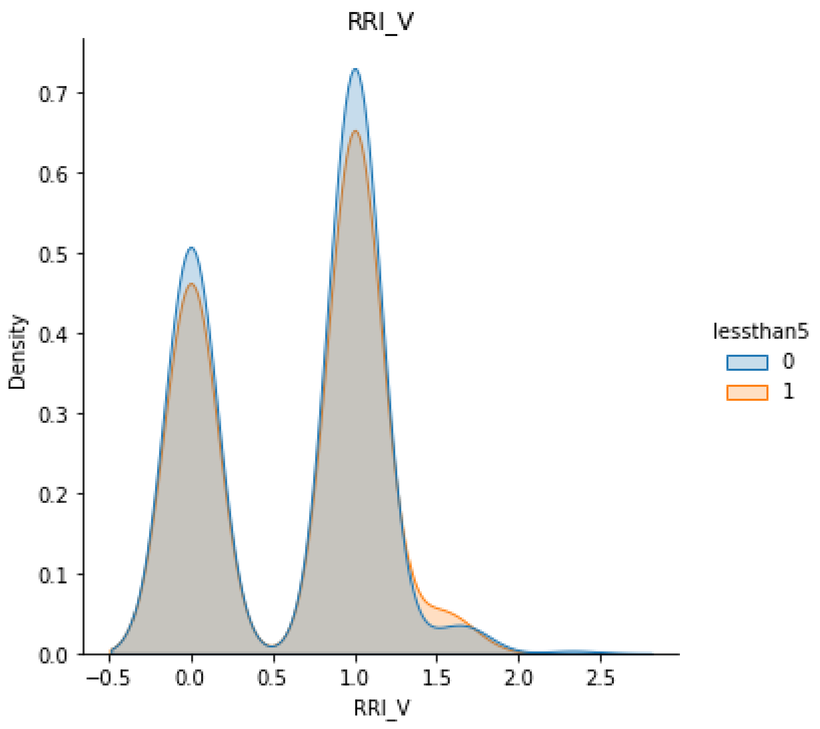

RRI_P
Density
RRI_P
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

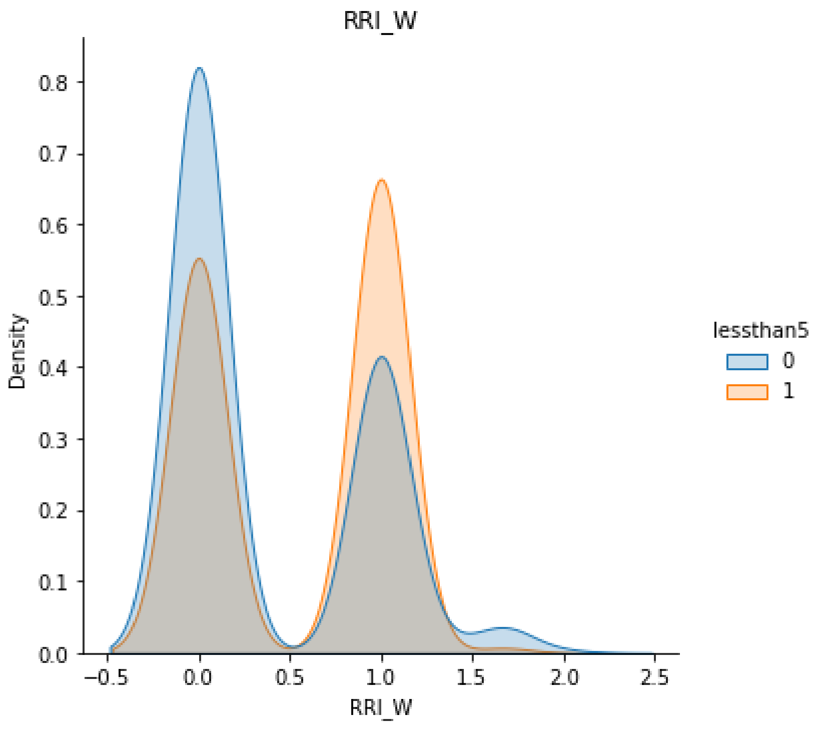

RRI_Q
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
RRI_Q
lessthan5
0
1

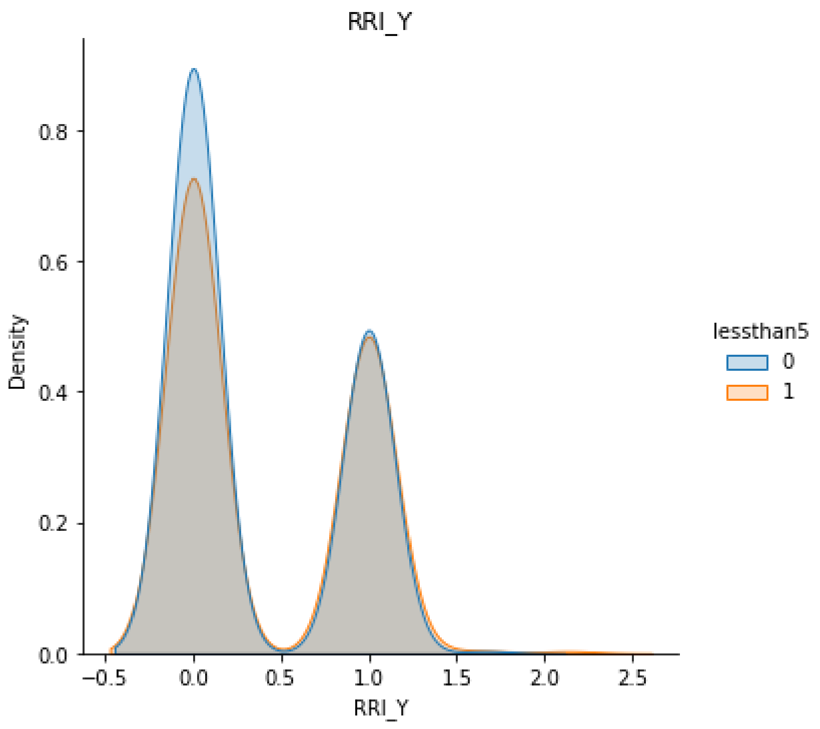

RRI_R
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
RRI_R
lessthan5
0
1

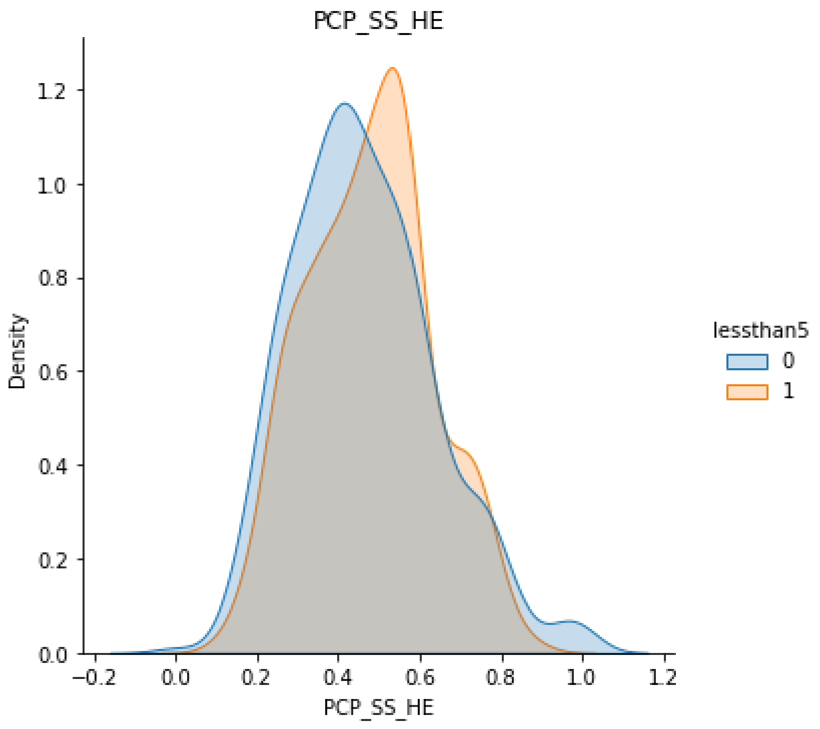

RRI_S
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
RRI_S
lessthan5
0
1

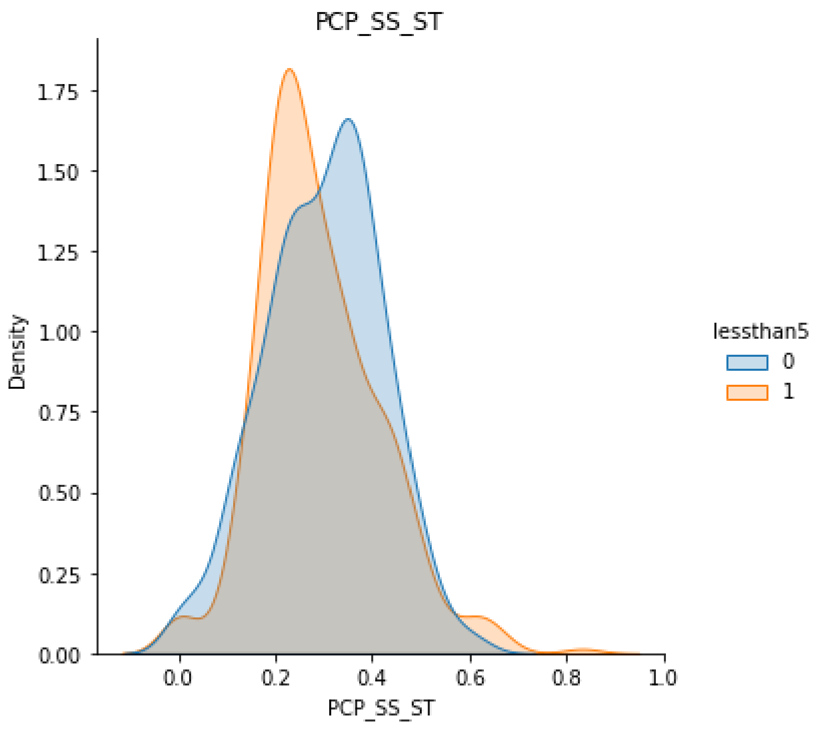

RRI_T
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
RRI_T
lessthan5
0
1

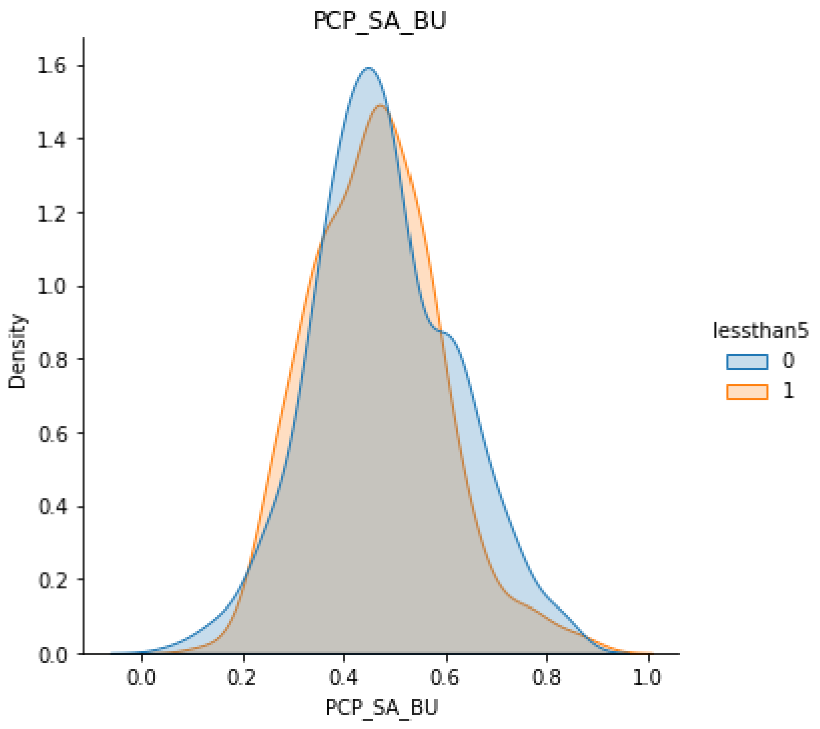

RRI_V
Density
RRI_V
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

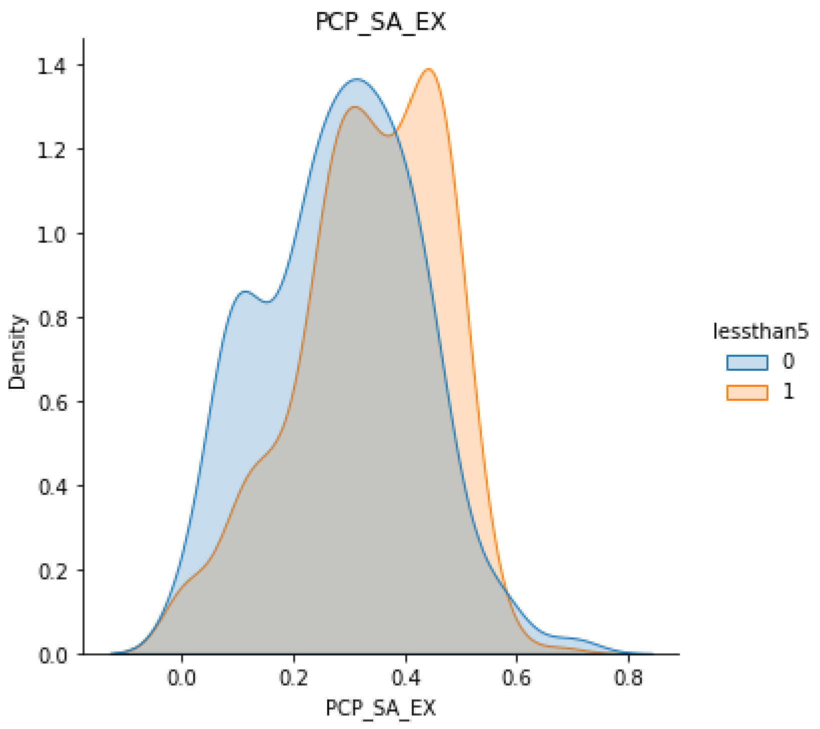

RRI_W
Density
RRI_W
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

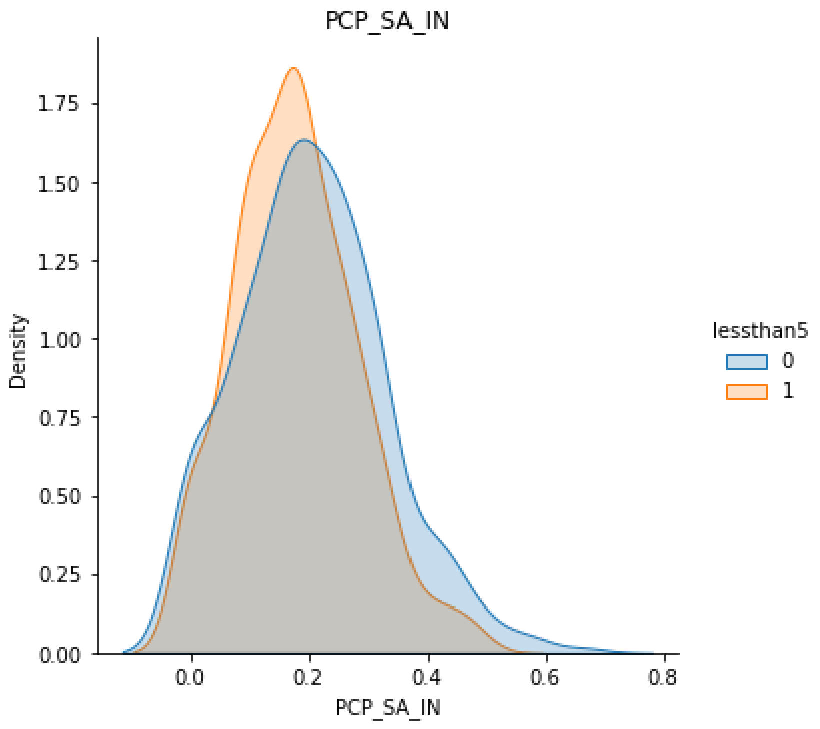

RRI_Y
Density
RRI_Y
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


PCP_SS_HE
Density
PCP_SS_HE
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

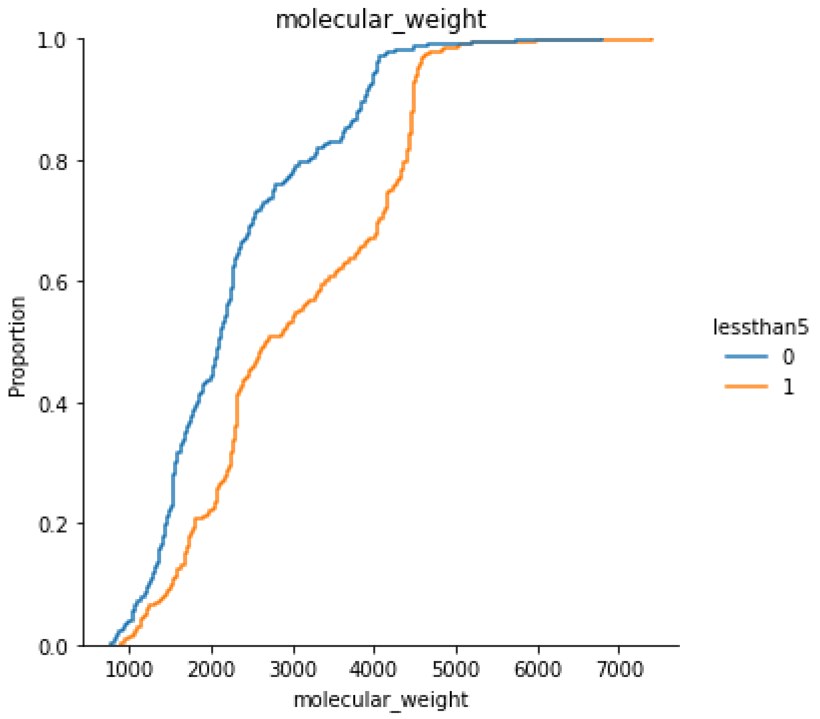

PCP_SS_ST
Density
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
PCP_SS_ST
lessthan5
0
1

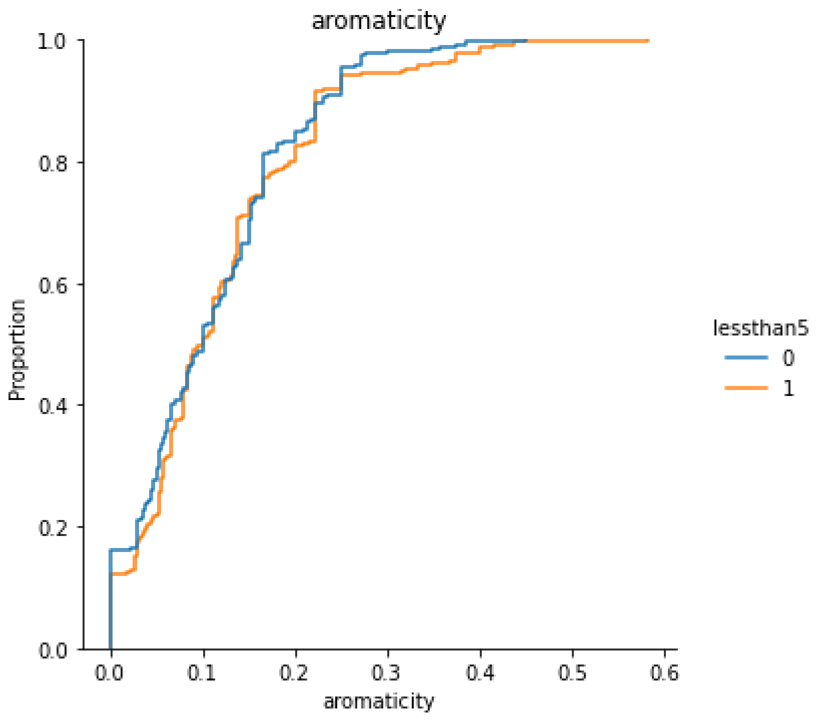

PCP_SA_BU
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PCP_SA_BU
lessthan5
0
1

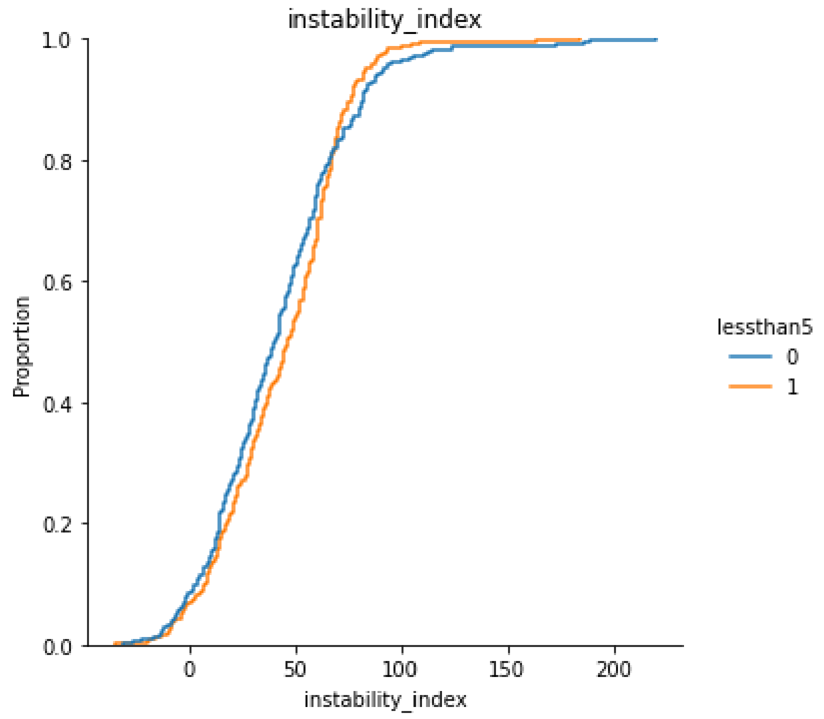

PCP_SA_EX
Density
PCP_SA_EX
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

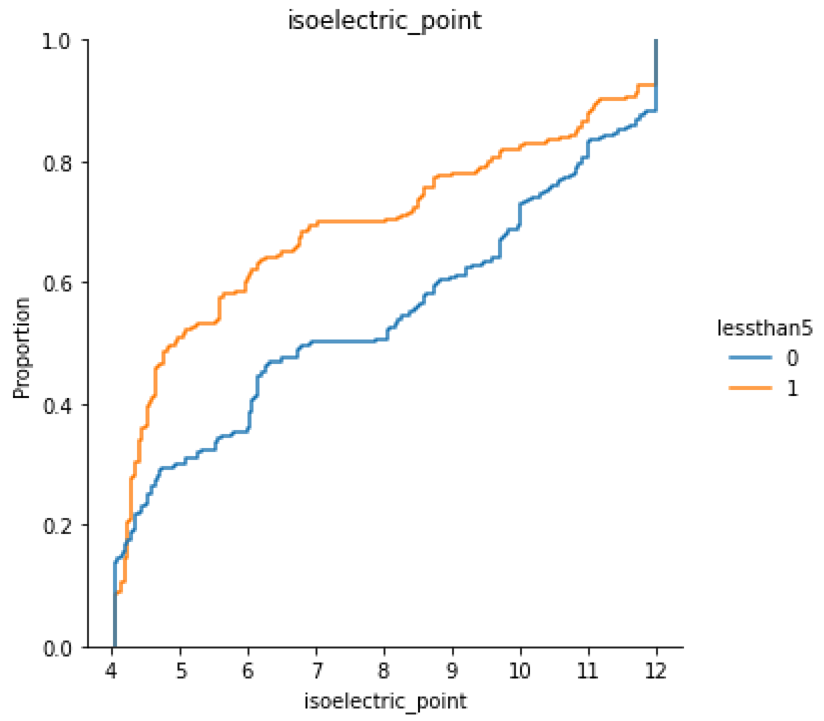

PCP_SA_IN
Density
PCP_SA_IN
lessthan5
0
1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.0
0.2
0.4
0.6
0.8

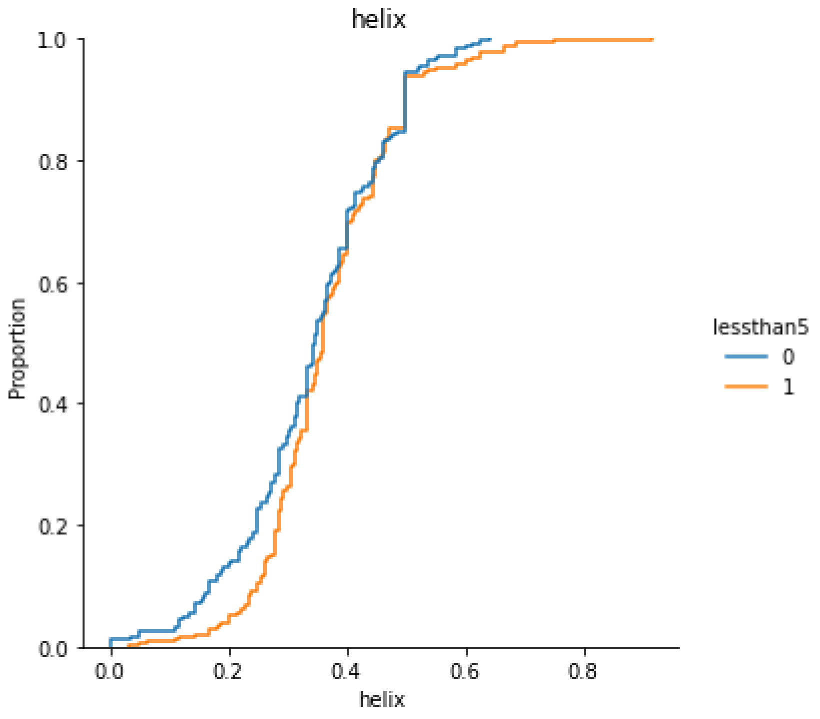
In [43]:
```
for feature in physicochem_prop.columns:
    sns.displot(x=physicochem_prop[feature], hue=df['lessthan5'], kind=
"ecdf")
    plt.title(feature)
    plt.show()
```

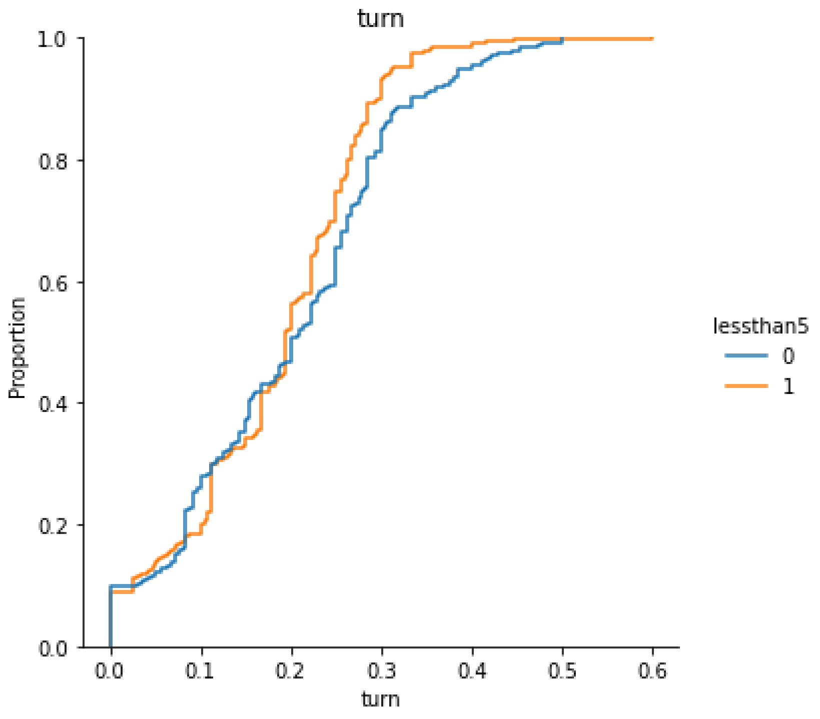

molecular_weight
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000
5000
6000
7000
molecular_weight
lessthan5
0
1

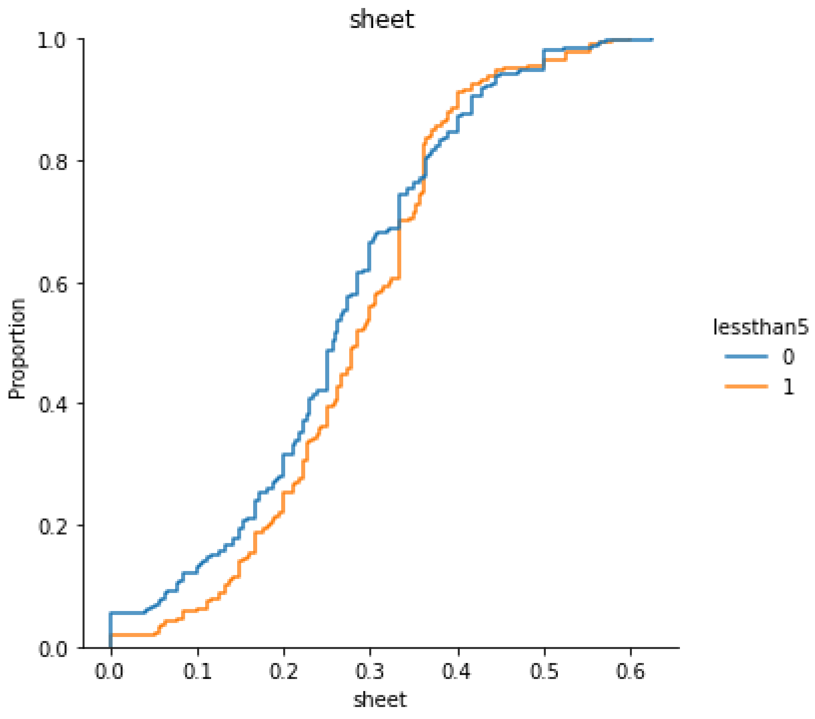

aromaticity
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
aromaticity
lessthan5
0
1

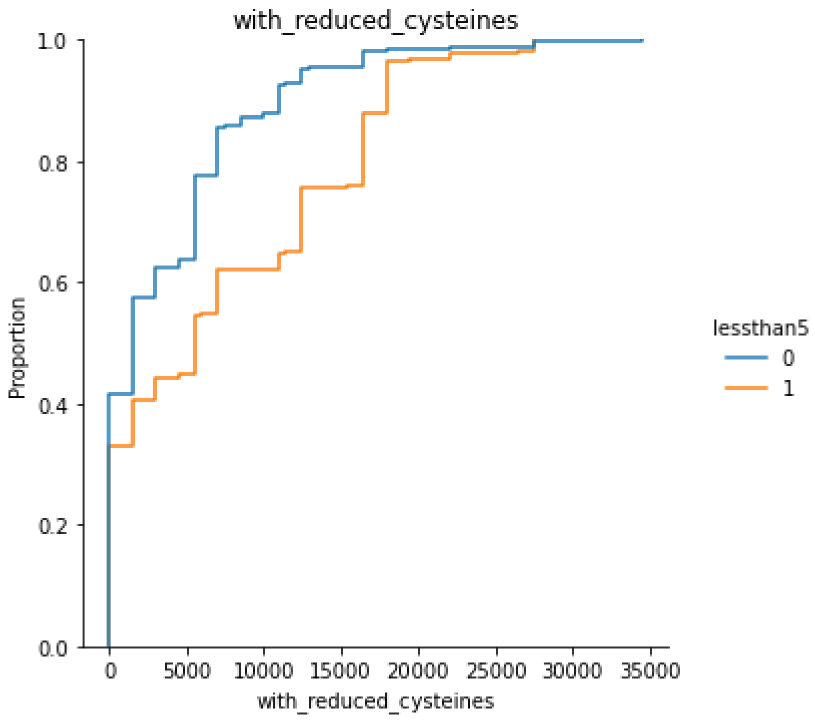

instability_index
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0
50
100
150
200
instability_index
lessthan5
0
1

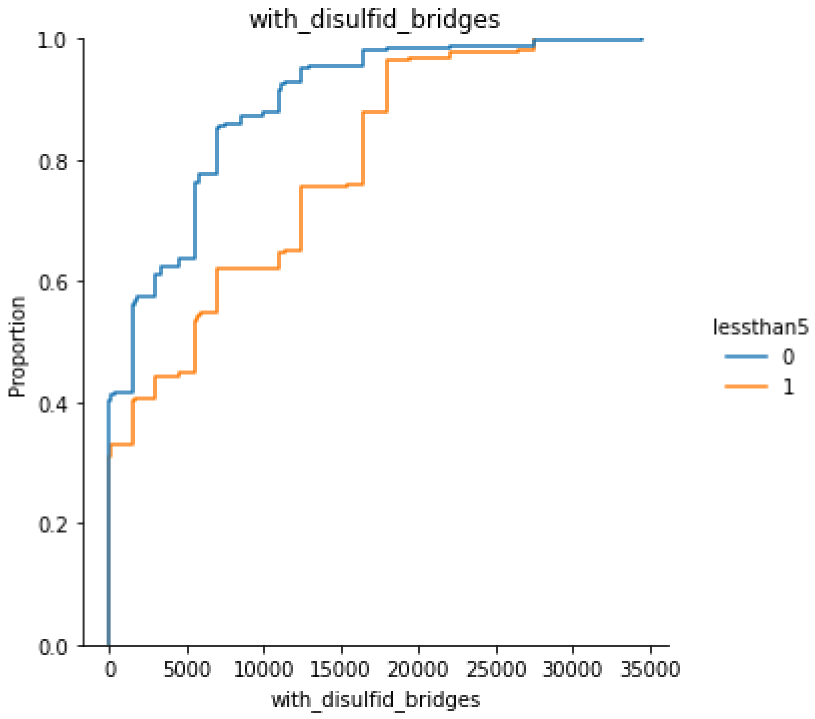

isoelectric_point
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
4
5
6
7
8
9
10
11
12
isoelectric_point
lessthan5
0
1

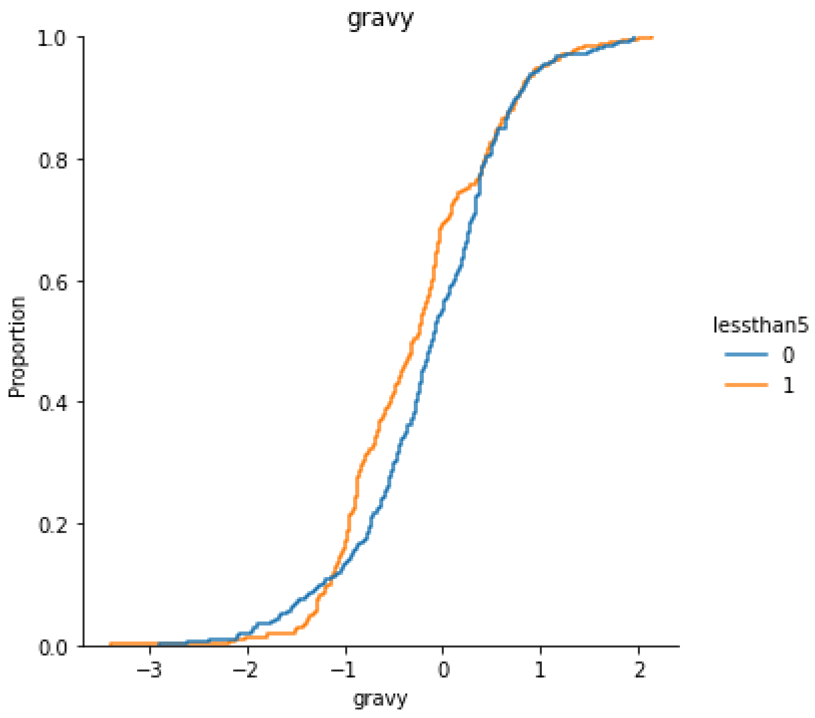

helix
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
helix
lessthan5
0
1

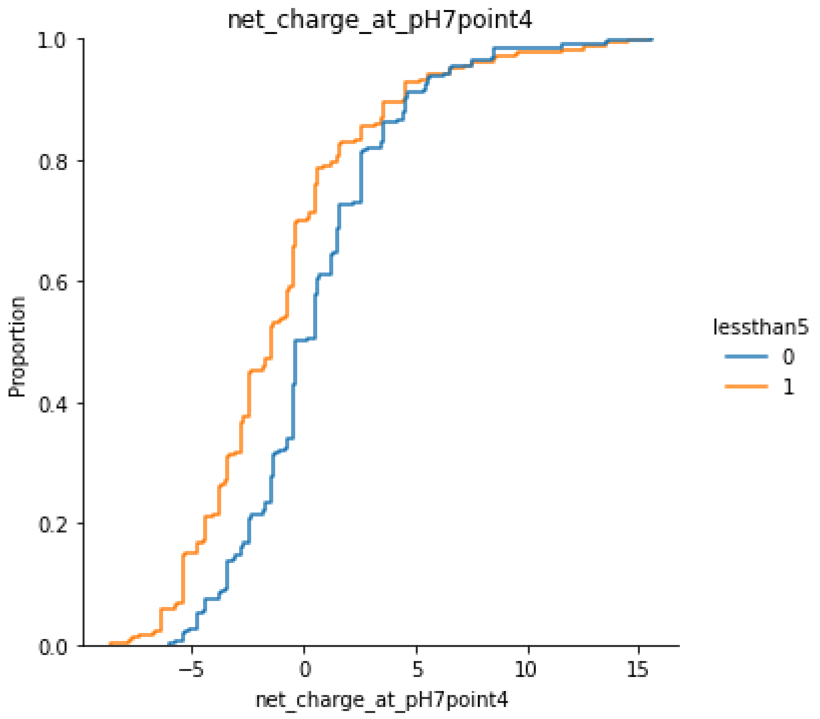

turn
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
turn
lessthan5
0
1

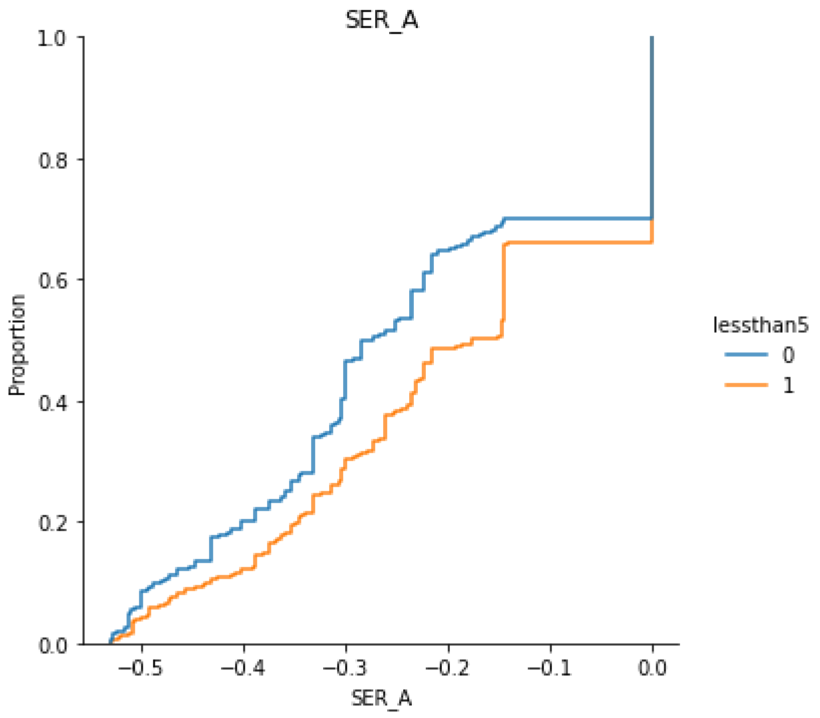

sheet
Proportion
sheet
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6

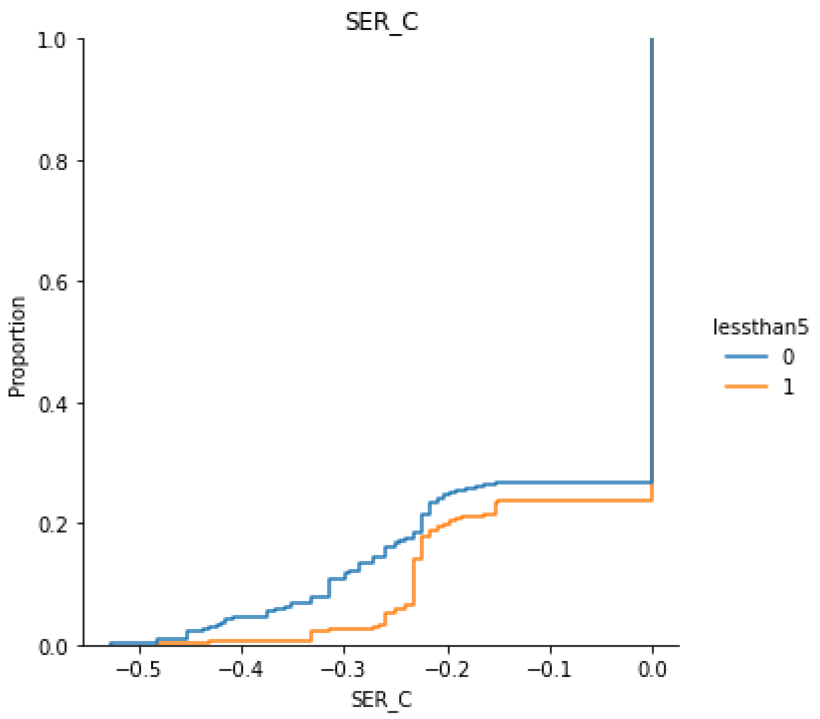

with_reduced_cysteines
Proportion
with_reduced_cysteines
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
25000
30000
35000

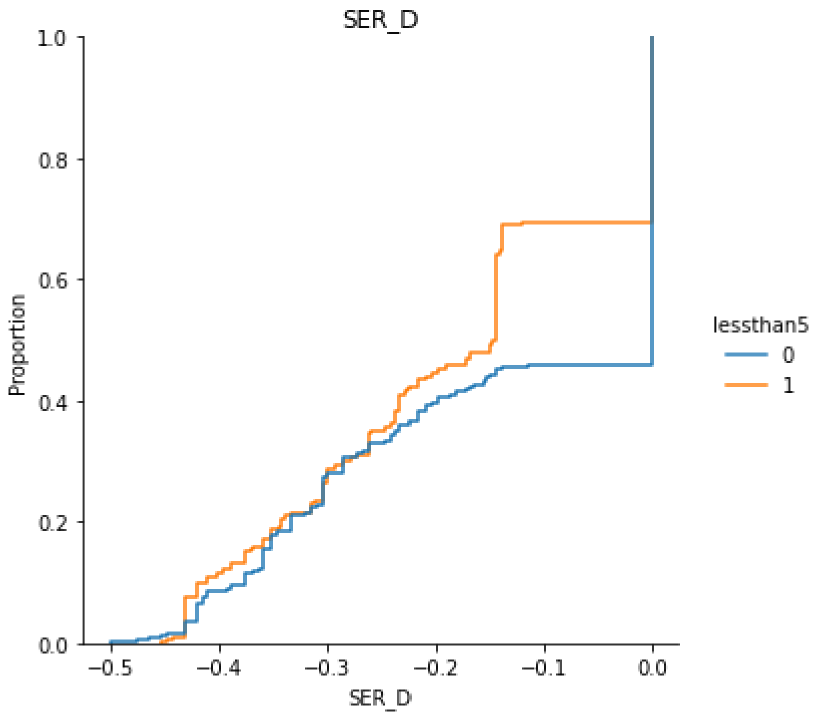

with_disulfid_bridges
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000
35000
with_disulfid_bridges
lessthan5
0
1

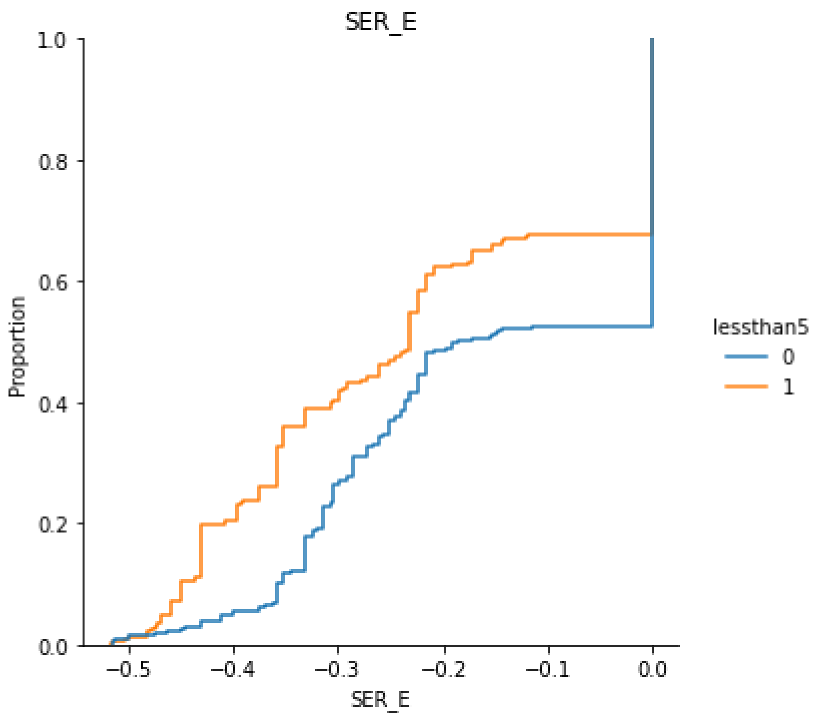

gravy
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
gravy
lessthan5
0
1

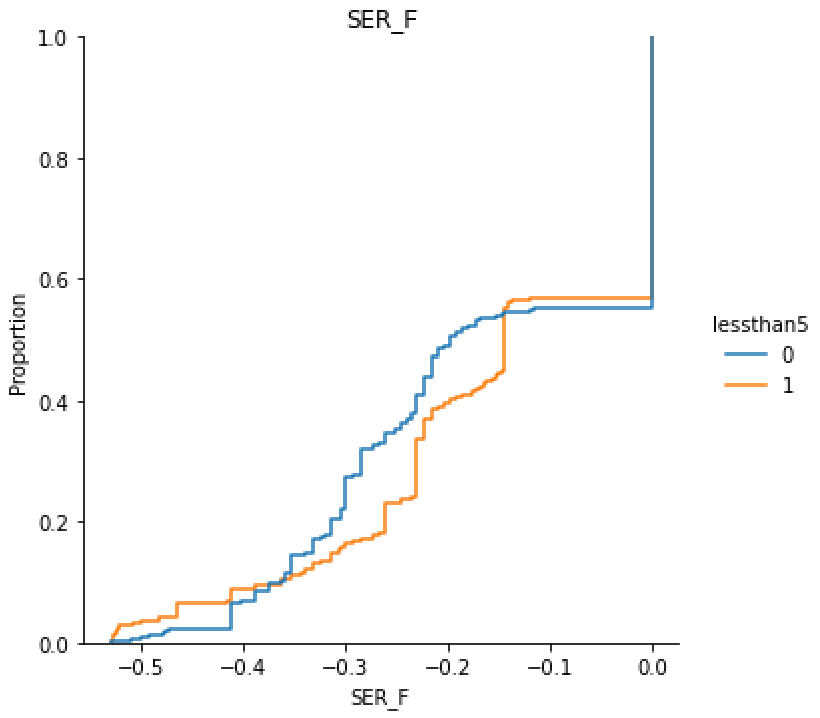

net_charge_at_pH7point4
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
−5
0
5
10
15
net_charge_at_pH7point4
lessthan5
0
1

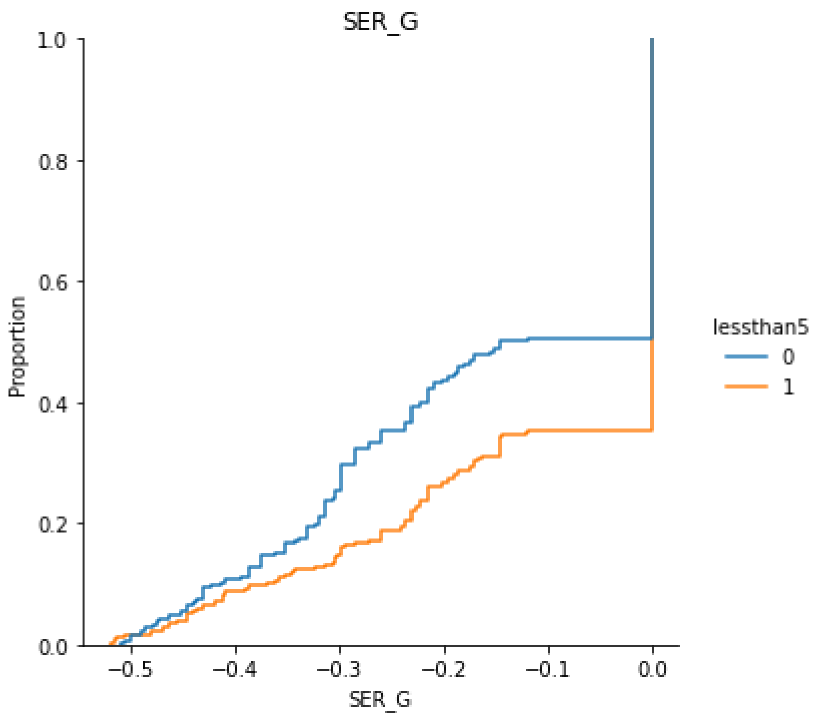

SER_A
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_A
lessthan5
0
1

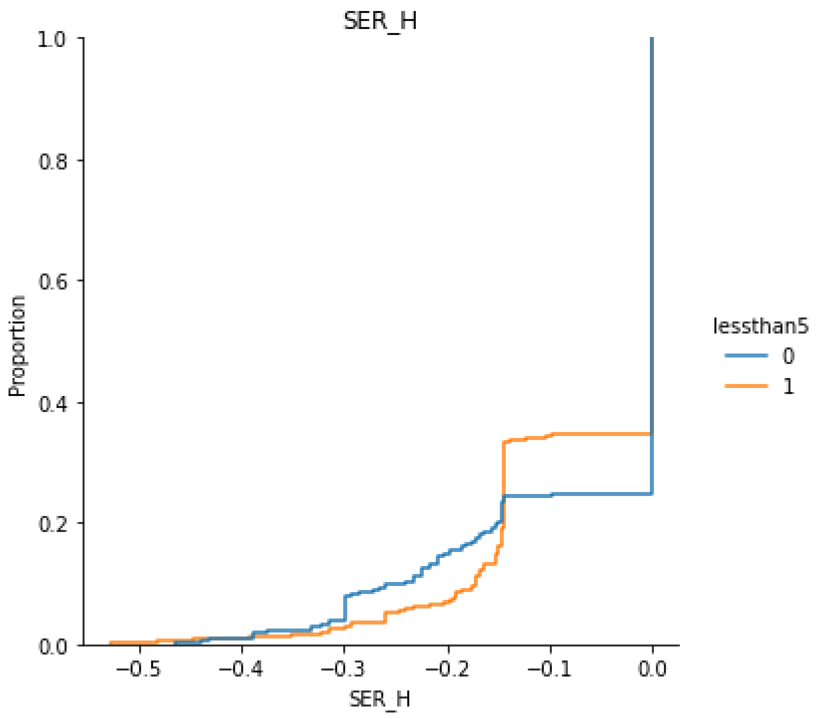

SER_C
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_C
lessthan5
0
1

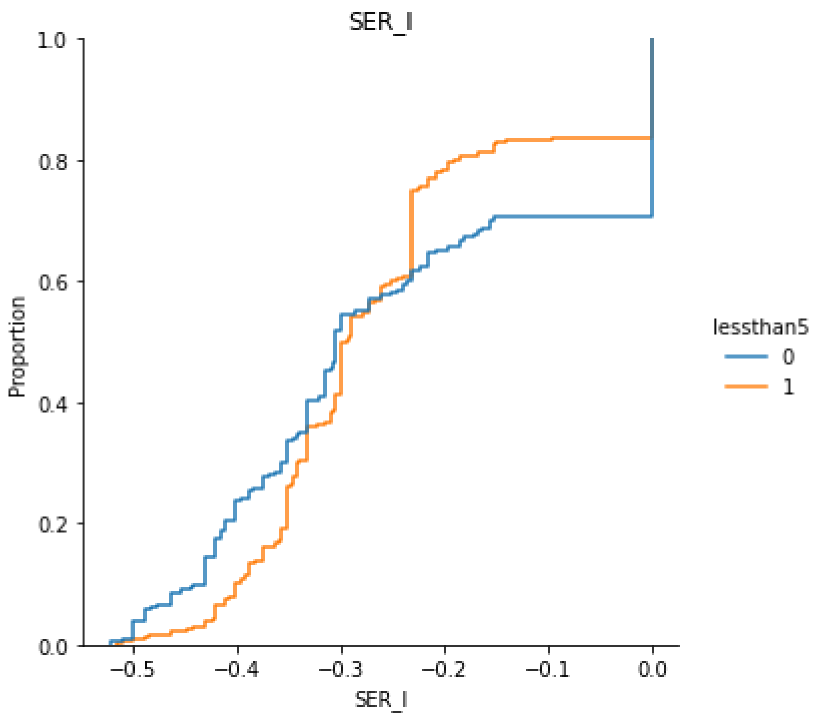

SER_D
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_D
lessthan5
0
1

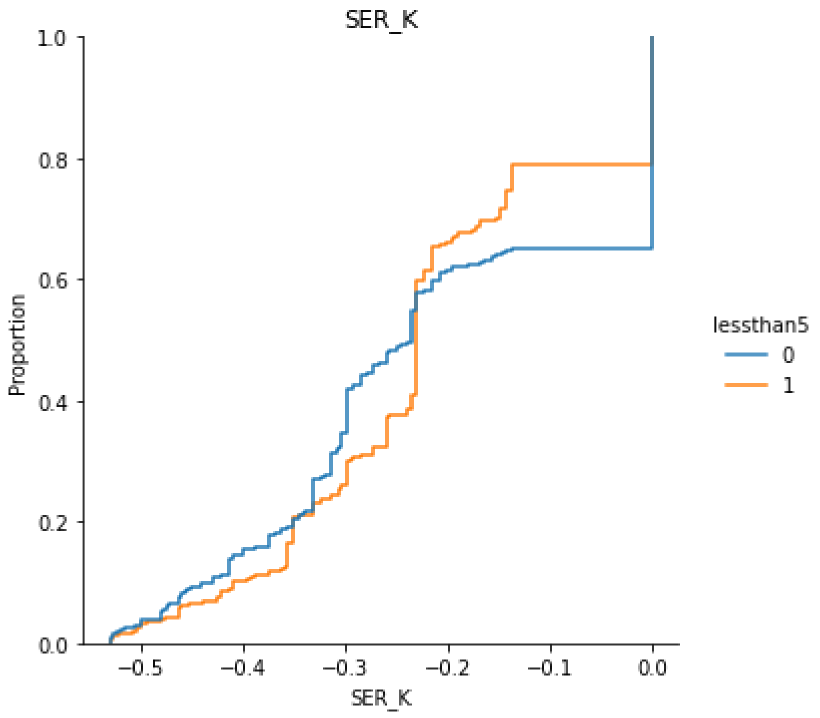

SER_E
Proportion
SER_E
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

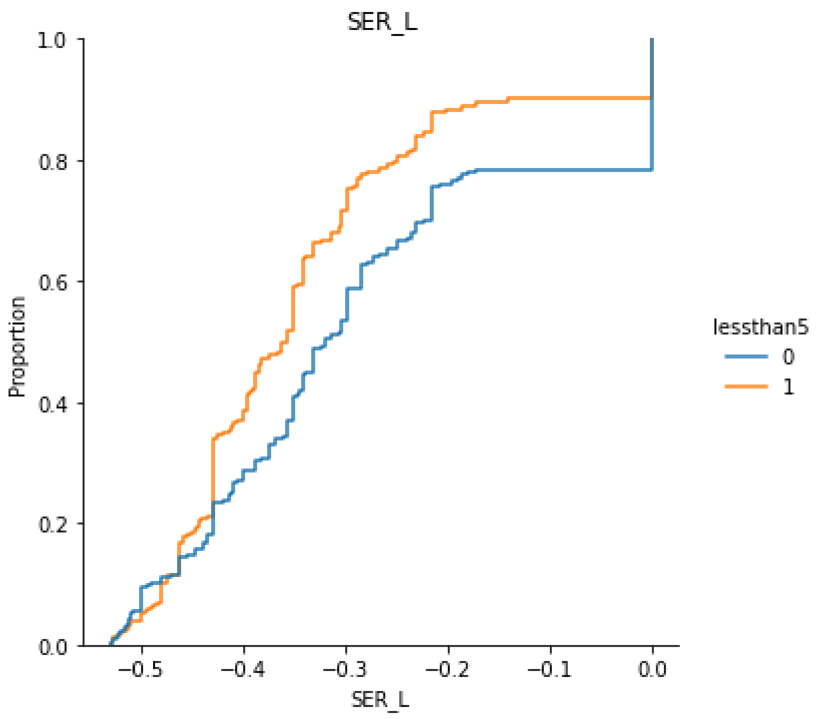

SER_F
Proportion
SER_F
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

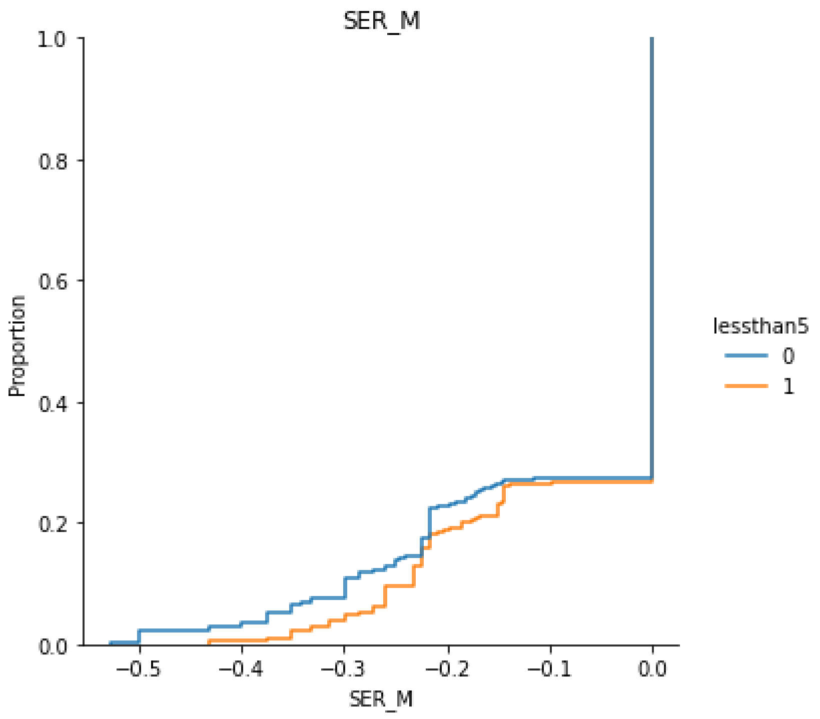

SER_G
Proportion
SER_G
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

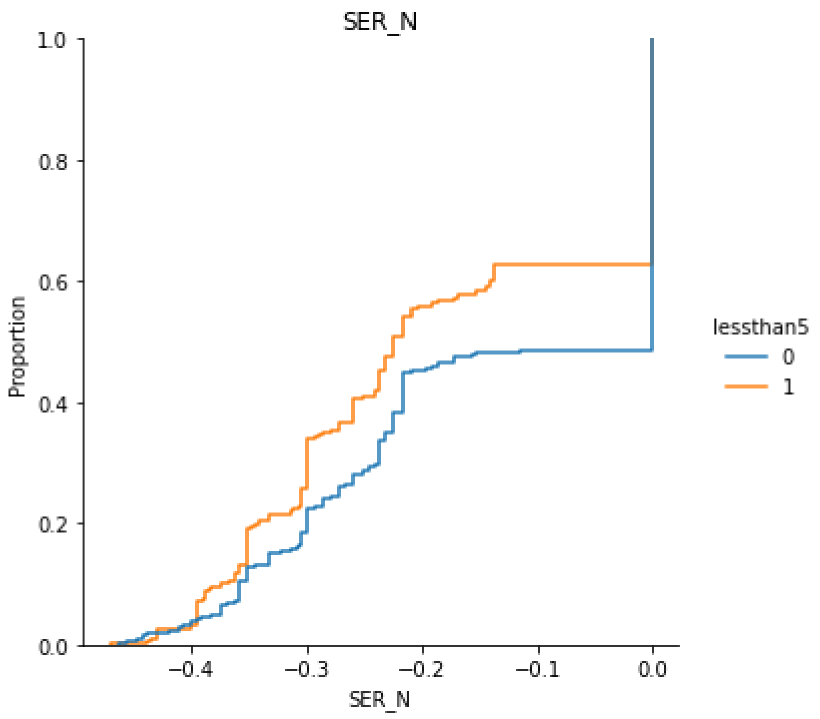

SER_H
Proportion
SER_H
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

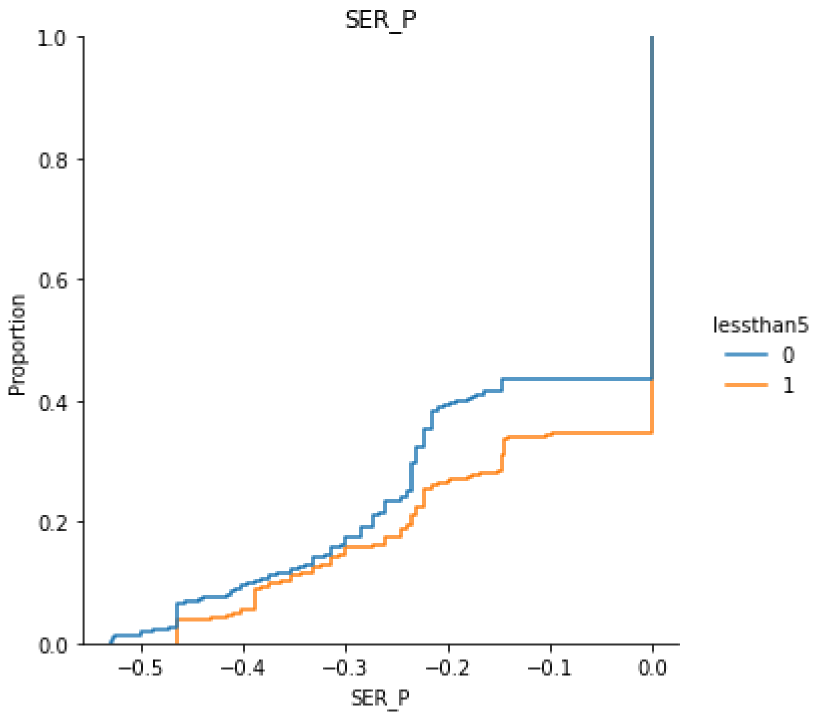

SER_I
Proportion
SER_I
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

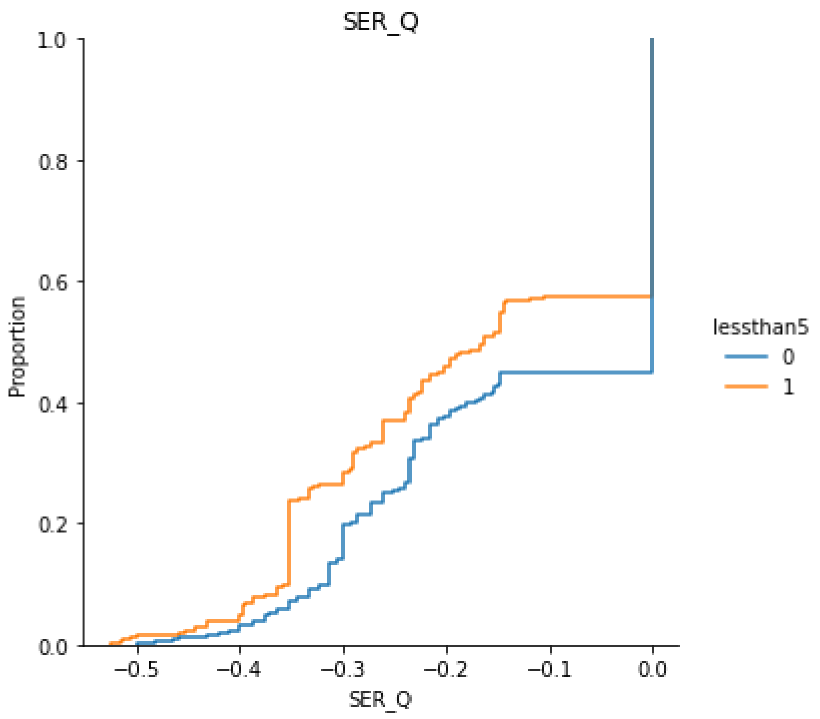

SER_K
Proportion
SER_K
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

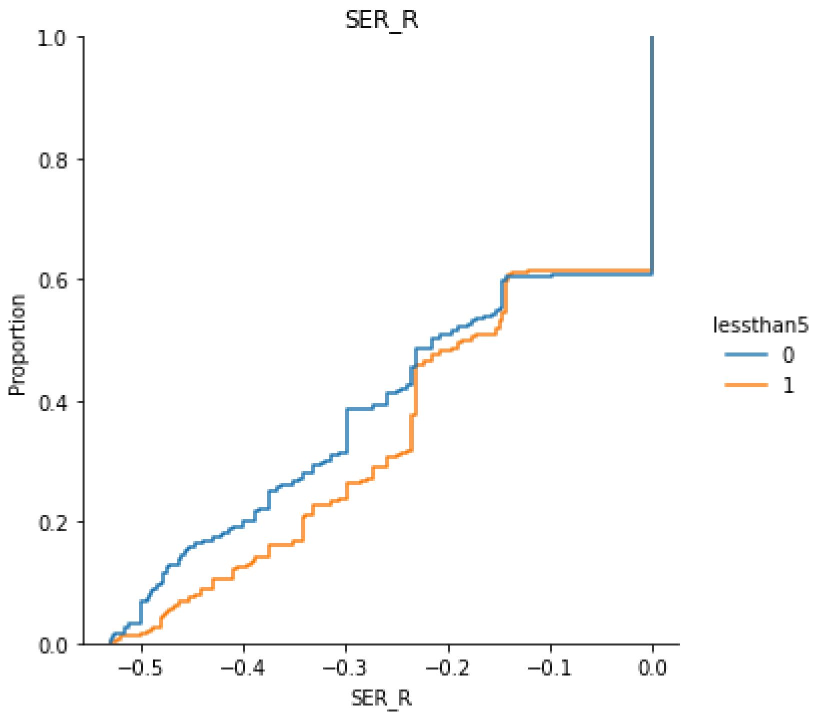

SER_L
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_L
lessthan5
0
1

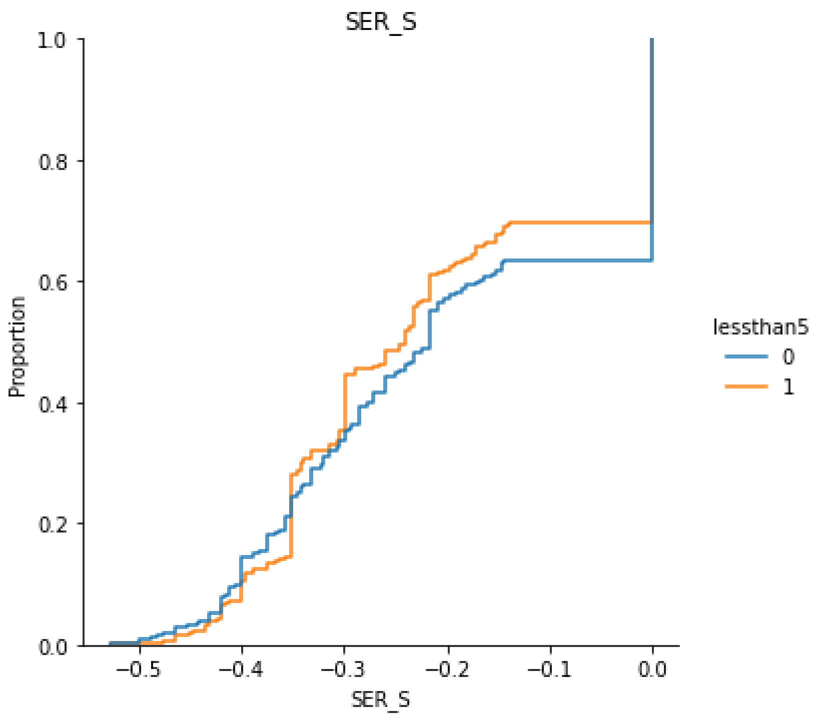

SER_M
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_M
lessthan5
0
1

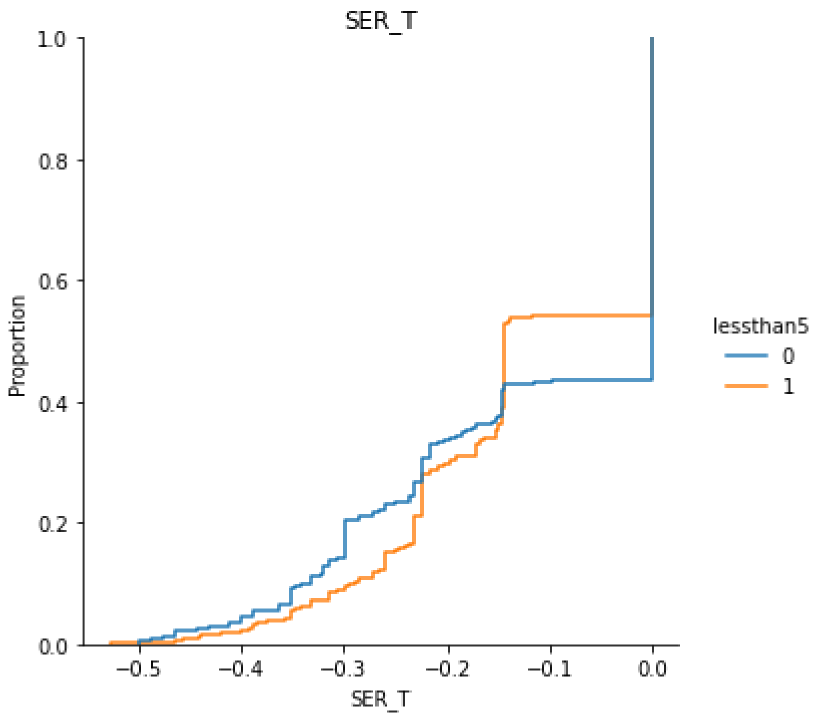

SER_N
Proportion
SER_N
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.4
−0.3
−0.2
−0.1
0.0

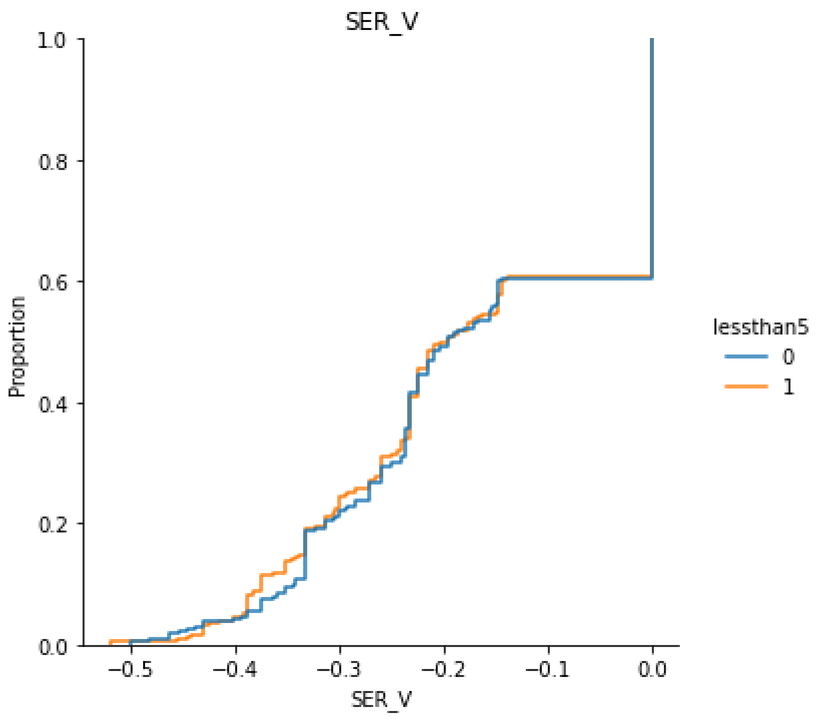

SER_P
Proportion
SER_P
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

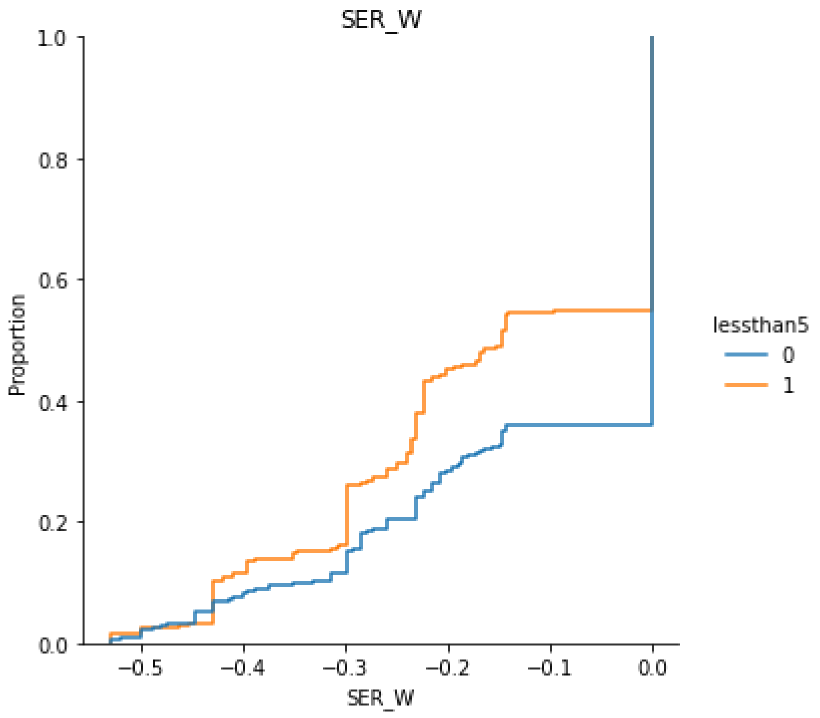

SER_Q
Proportion
SER_Q
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

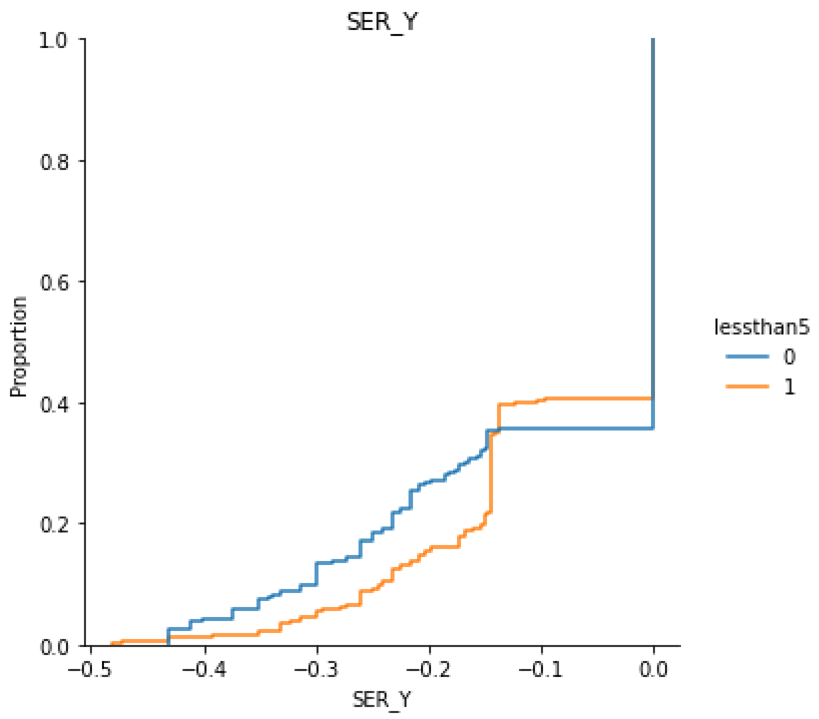

SER_R
Proportion
SER_R
lessthan5
0
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

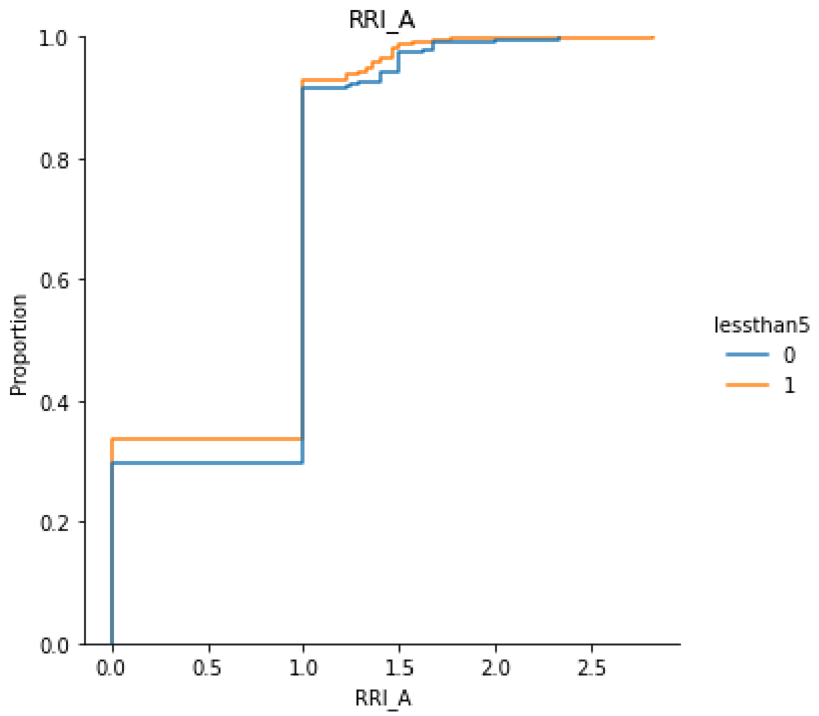

SER_S
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_S
lessthan5
0
1

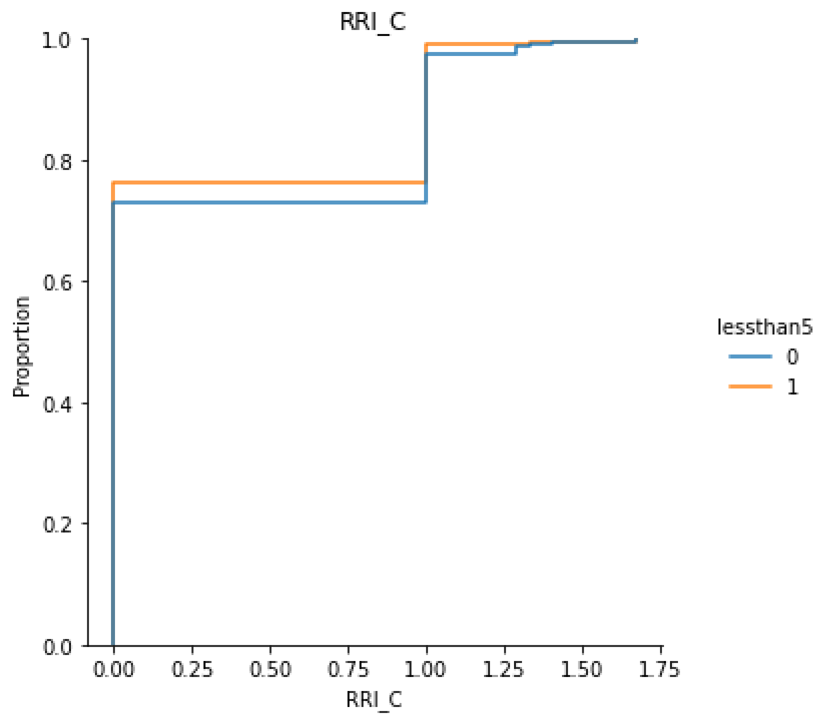

SER_T
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_T
lessthan5
0
1

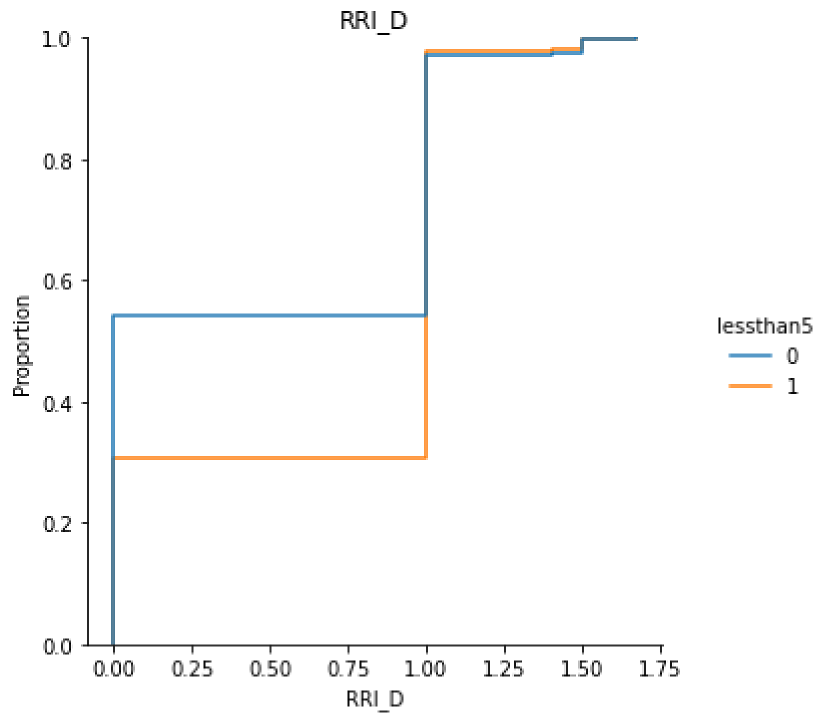

SER_V
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_V
lessthan5
0
1

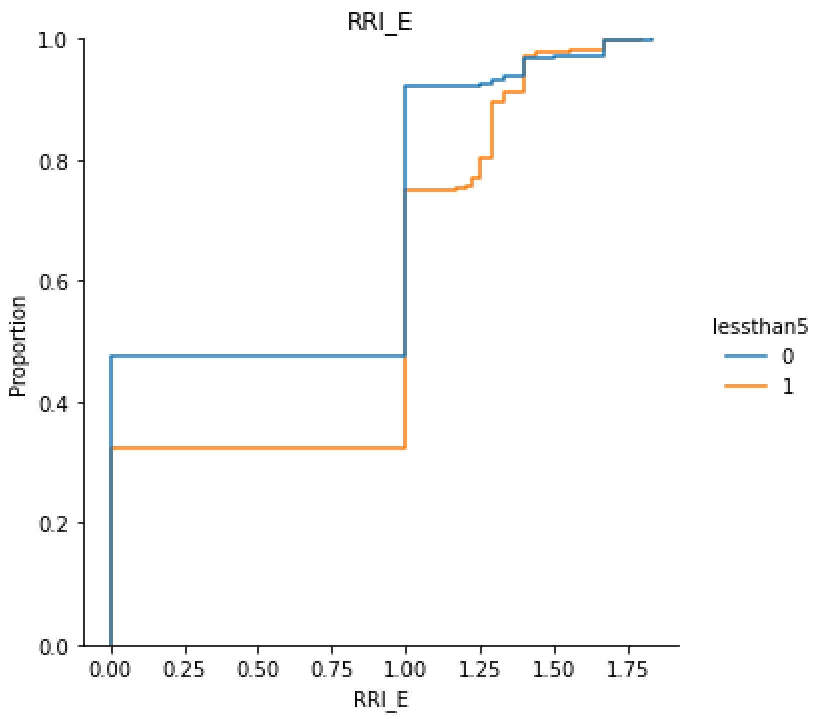

SER_W
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_W
lessthan5
0
1

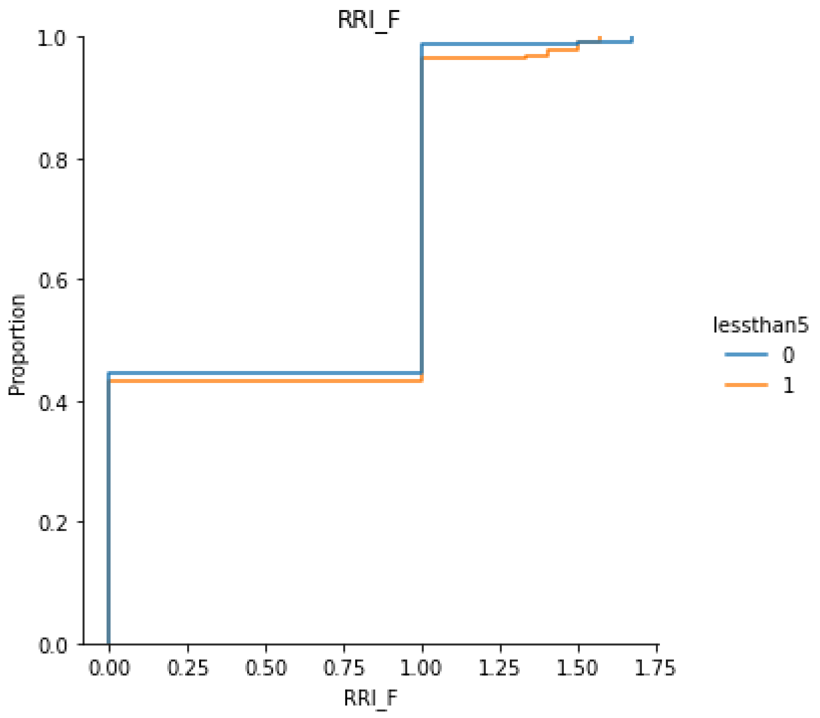

SER_Y
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_Y
lessthan5
0
1

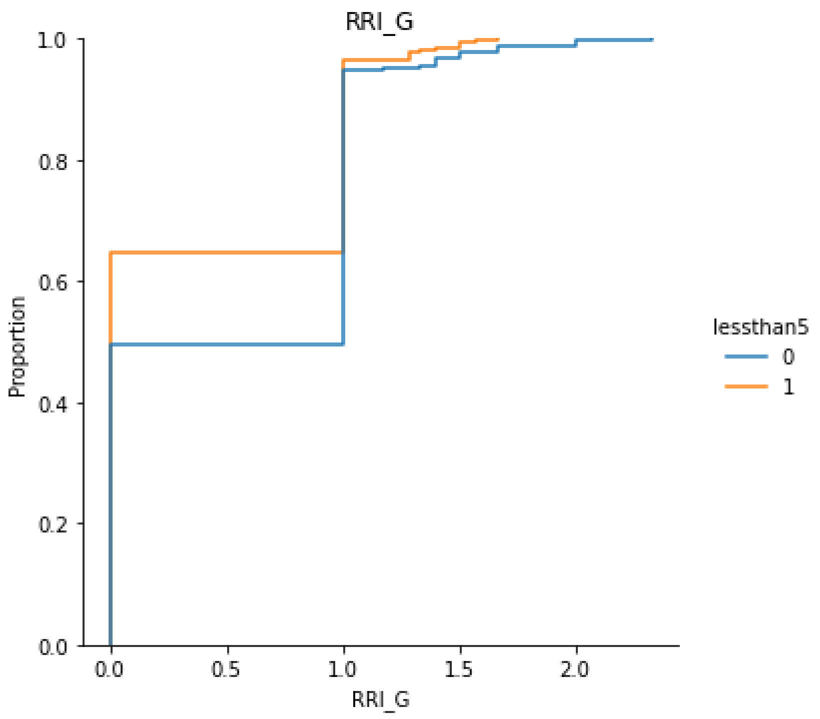

RRI_A
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.0
0.5
1.0
1.5
2.0
2.5
RRI_A
lessthan5
0
1

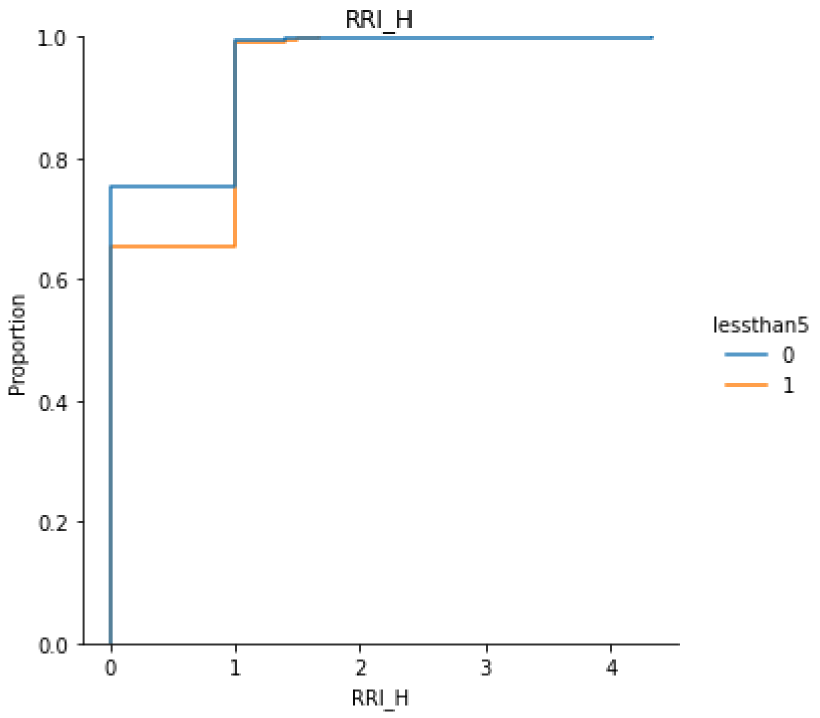

RRI_C
Proportion
RRI_C
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

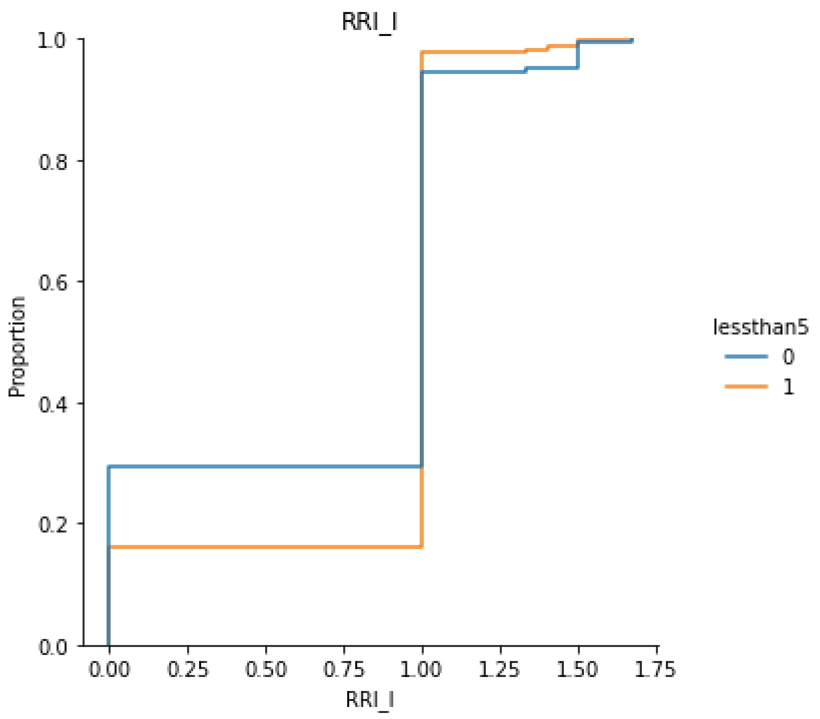

RRI_D
Proportion
RRI_D
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

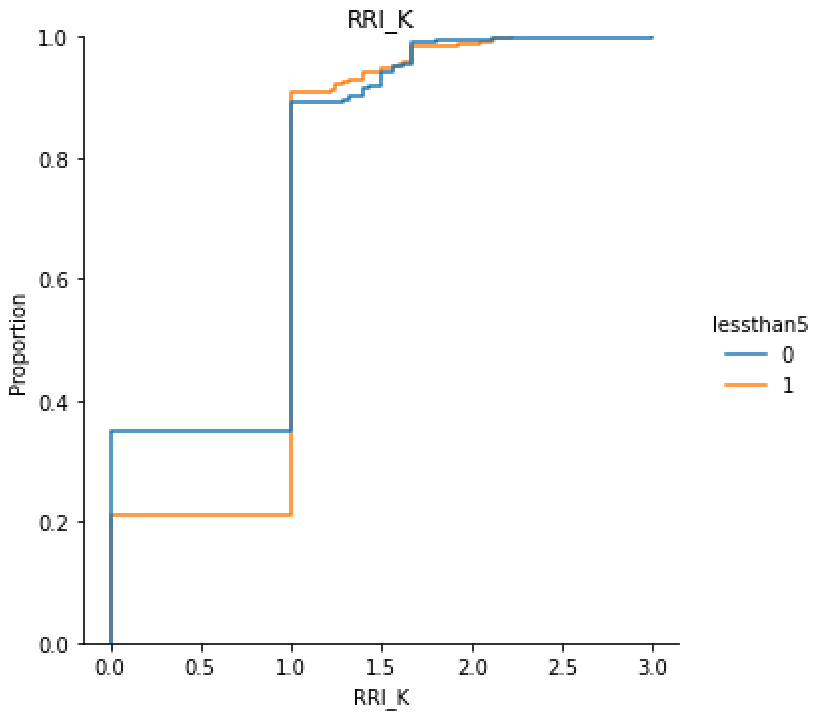

RRI_E
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
RRI_E
lessthan5
0
1

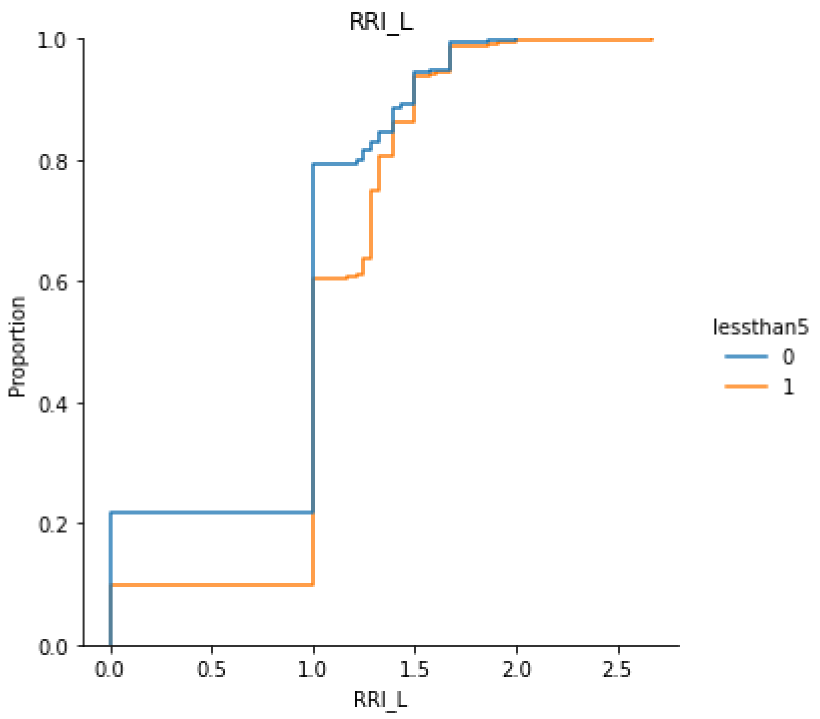

RRI_F
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
RRI_F
lessthan5
0
1

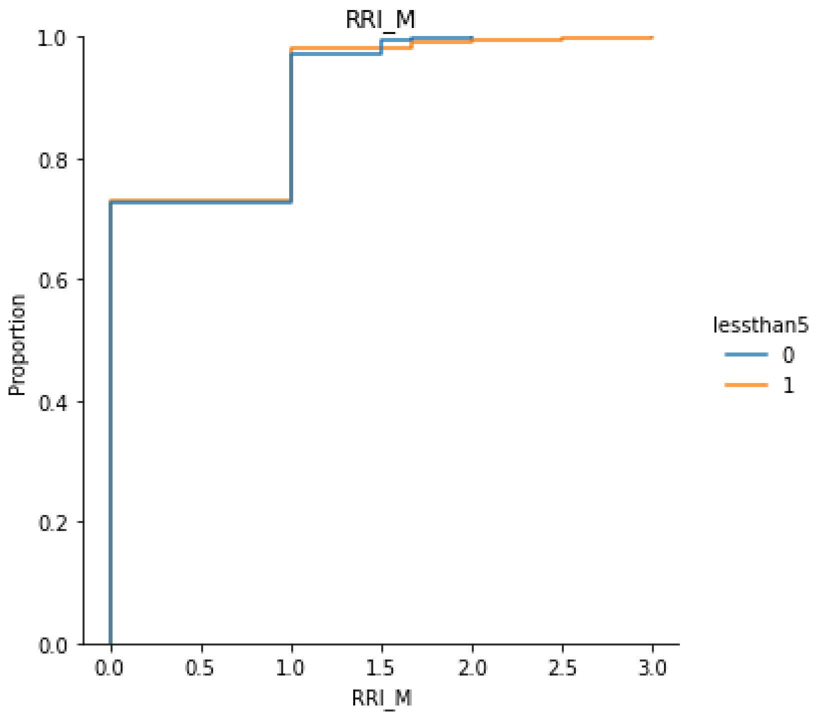

RRI_G
Proportion
RRI_G
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

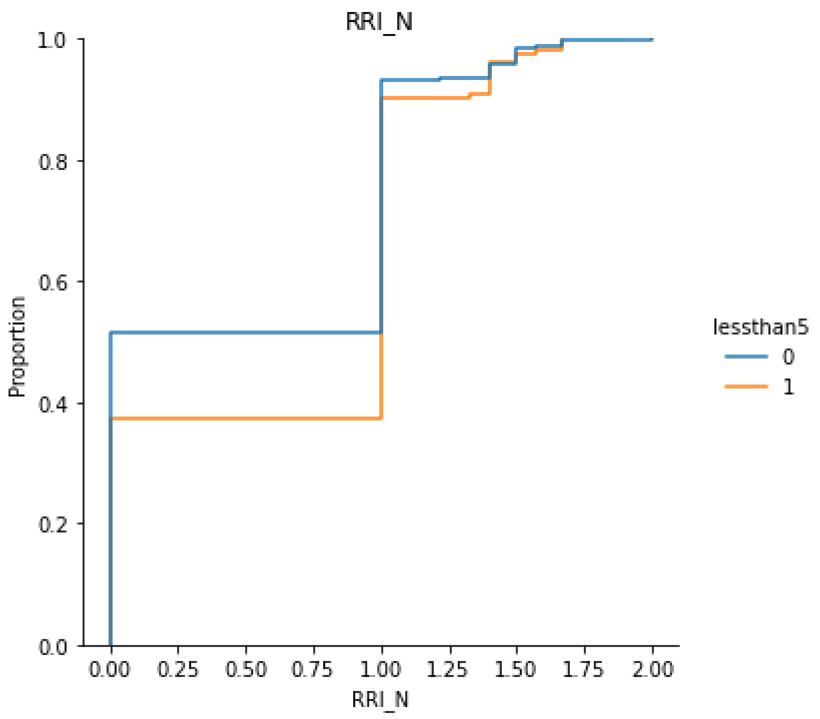

RRI_H
Proportion
RRI_H
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

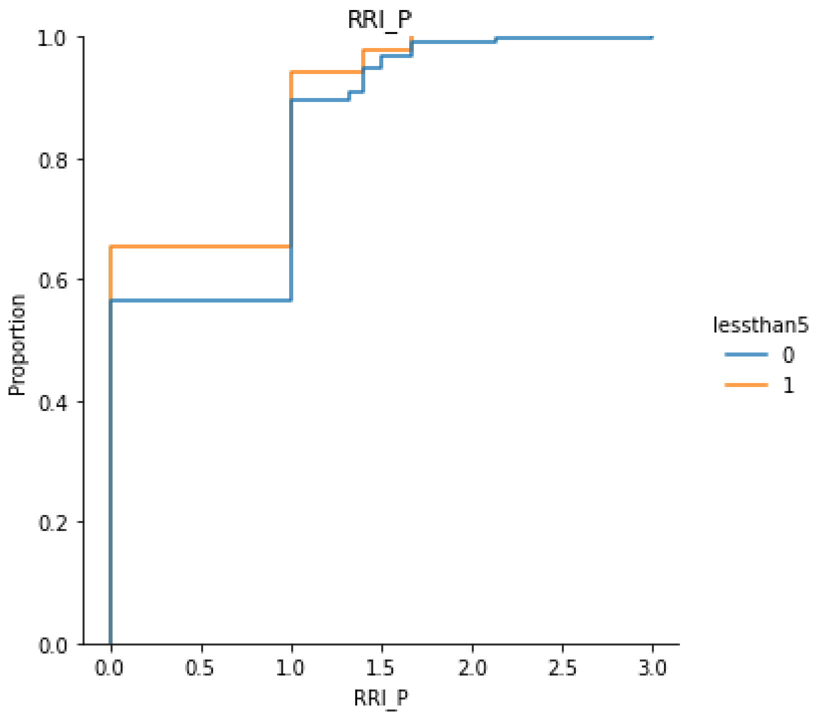

RRI_I
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
RRI_I
lessthan5
0
1

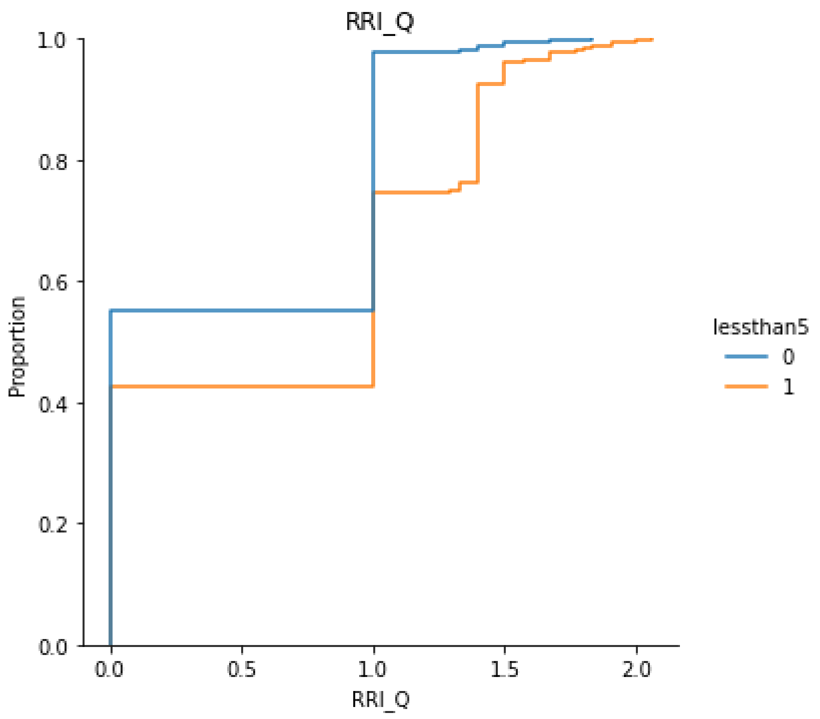

RRI_K
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RRI_K
lessthan5
0
1

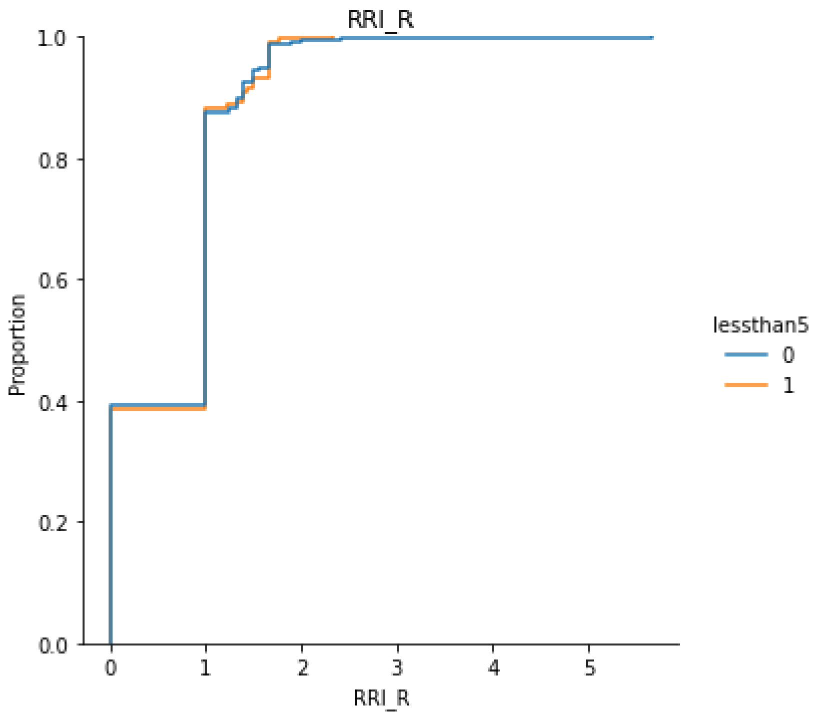

RRI_L
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
RRI_L
lessthan5
0
1

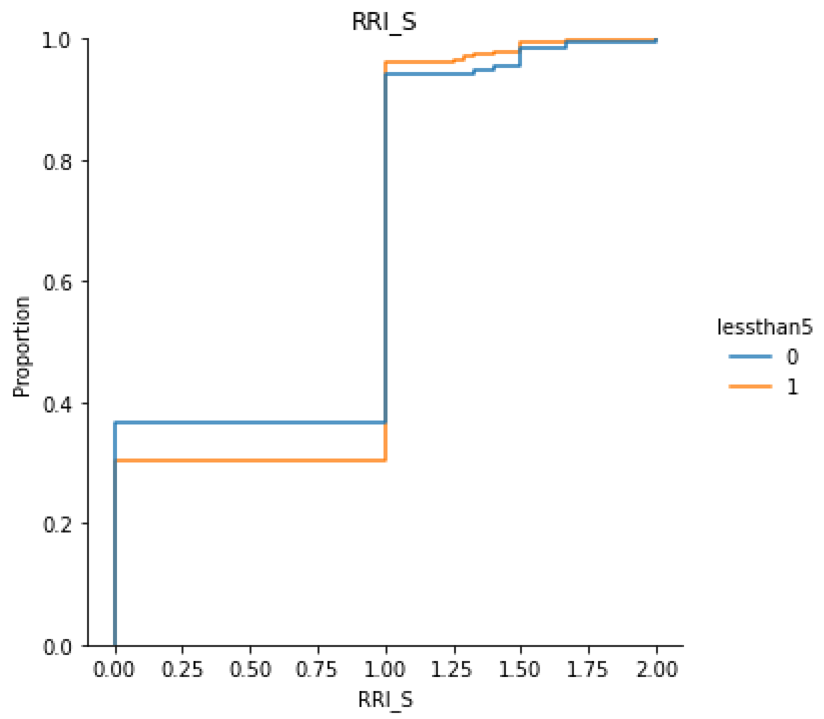

RRI_M
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RRI_M
lessthan5
0
1

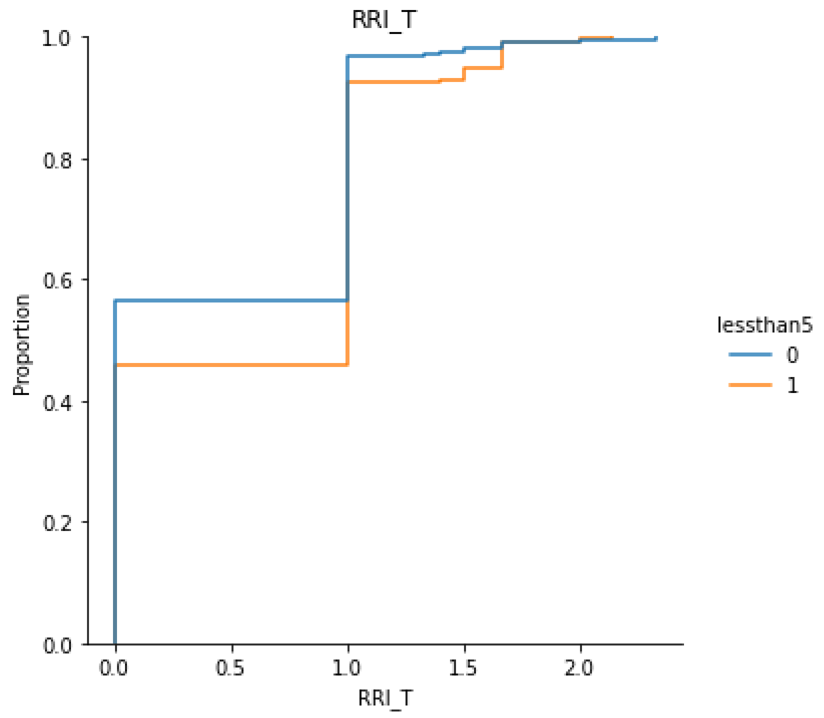

RRI_N
Proportion
RRI_N
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

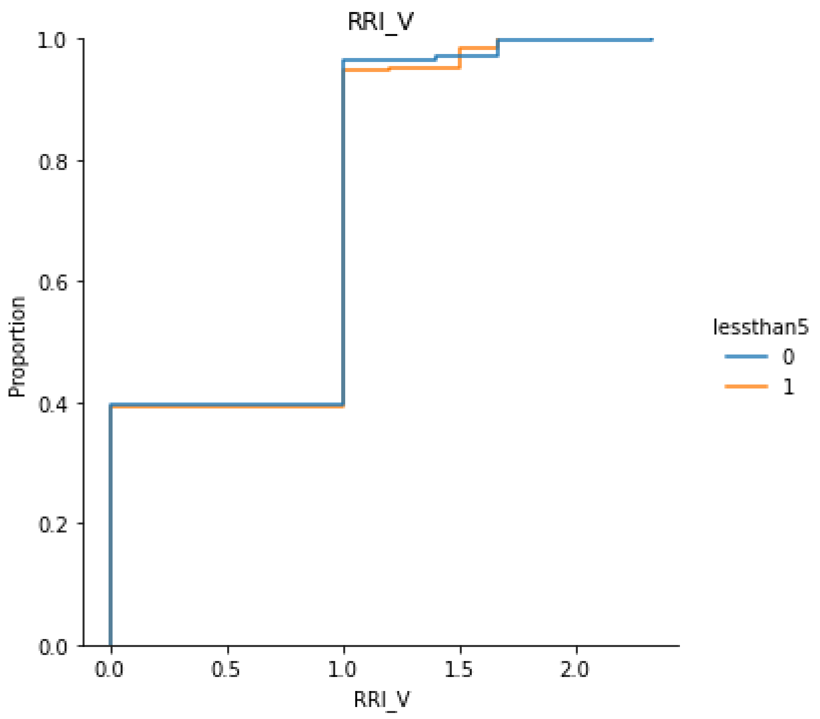

RRI_P
Proportion
RRI_P
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

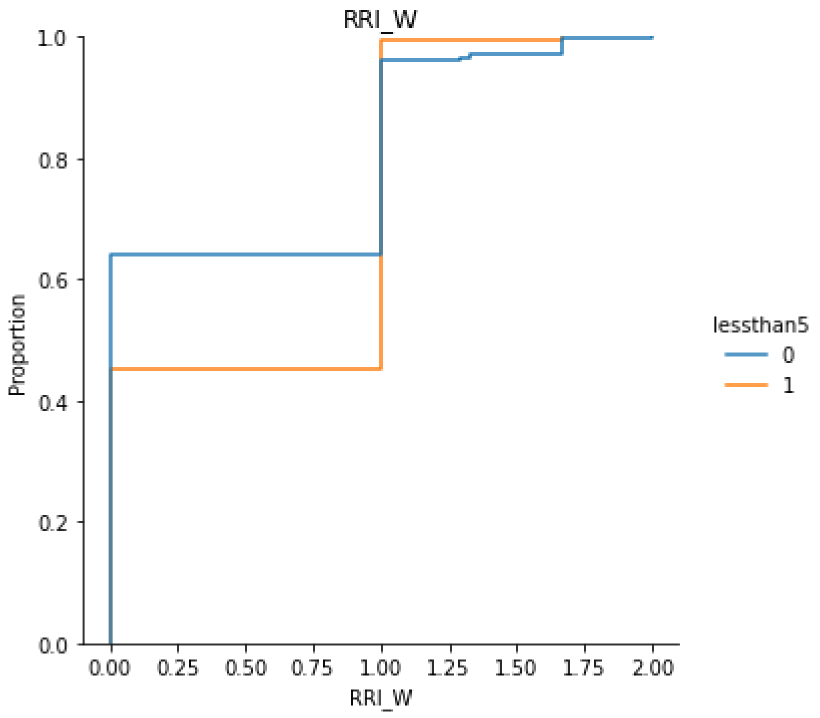

RRI_Q
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
RRI_Q
lessthan5
0
1

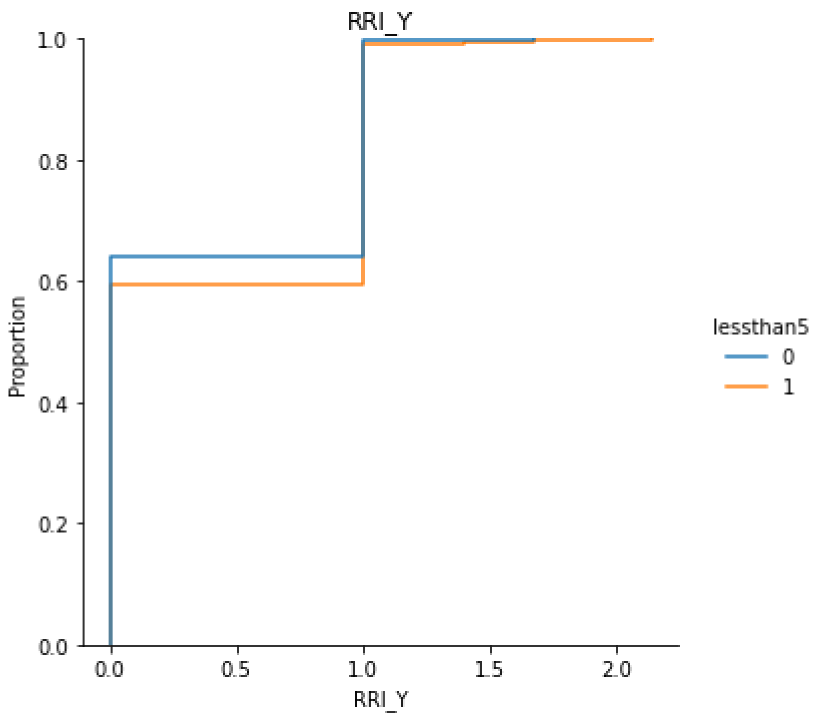

RRI_R
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
RRI_R
lessthan5
0
1

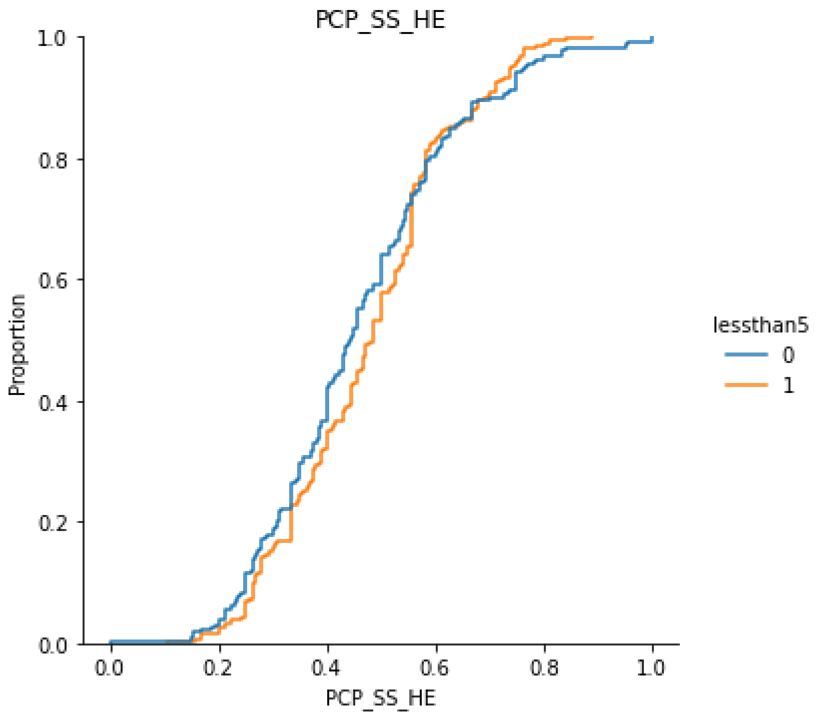

RRI_S
Proportion
RRI_S
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

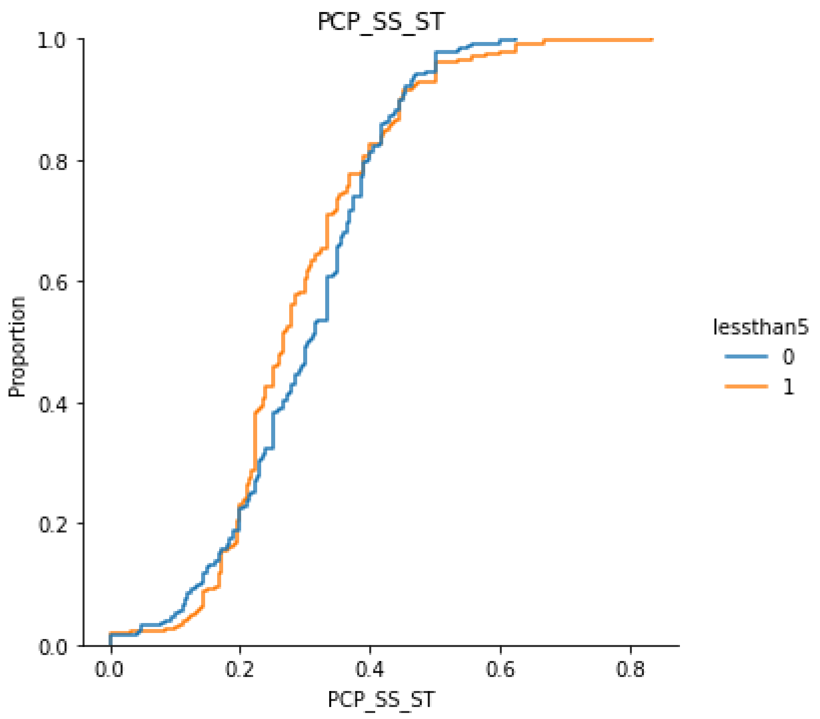

RRI_T
Proportion
RRI_T
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

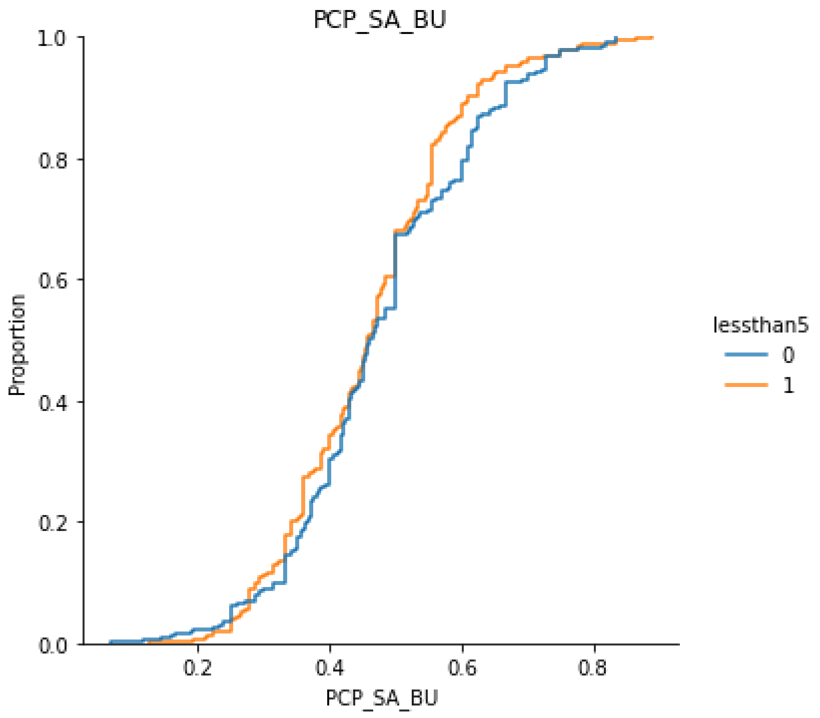

RRI_V
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.0
0.5
1.0
1.5
2.0
RRI_V
lessthan5
0
1

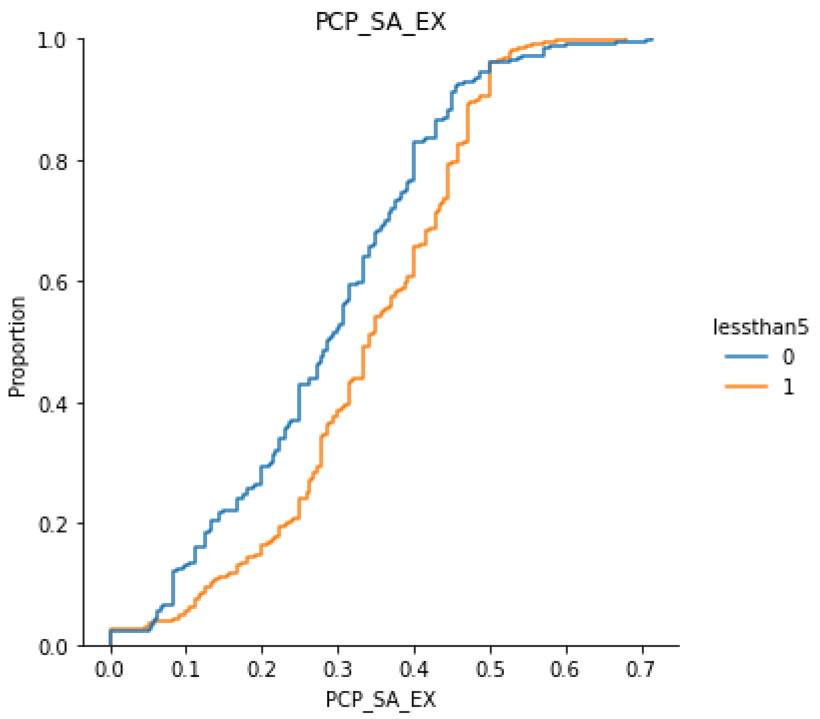

RRI_W
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
RRI_W
lessthan5
0
1

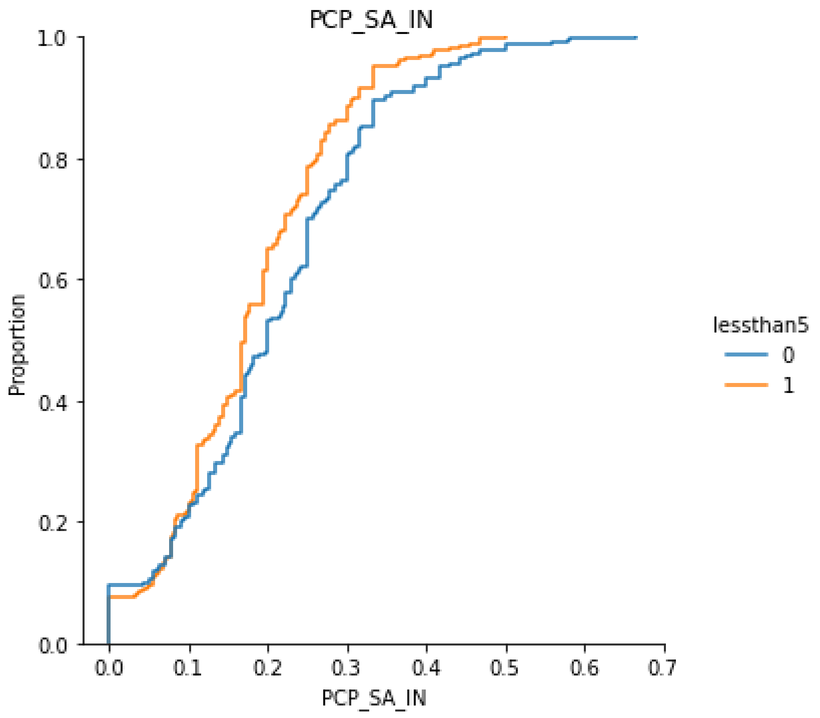

RRI_Y
Proportion
RRI_Y
lessthan5
0
1


PCP_SS_HE
Proportion
PCP_SS_HE
lessthan5
0
1

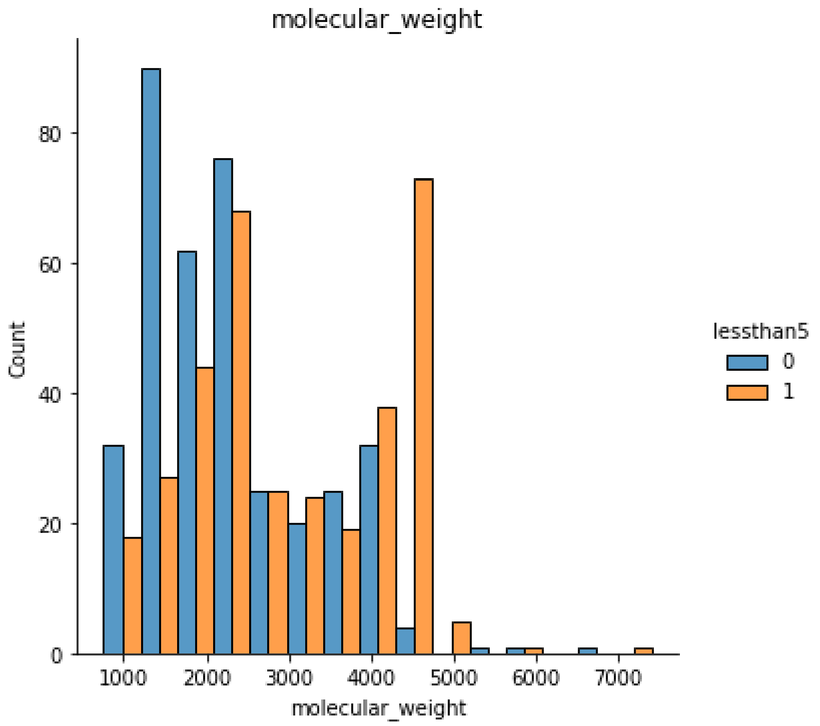

PCP_SS_ST
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.0
0.2
0.4
0.6
0.8
PCP_SS_ST
lessthan5
0
1

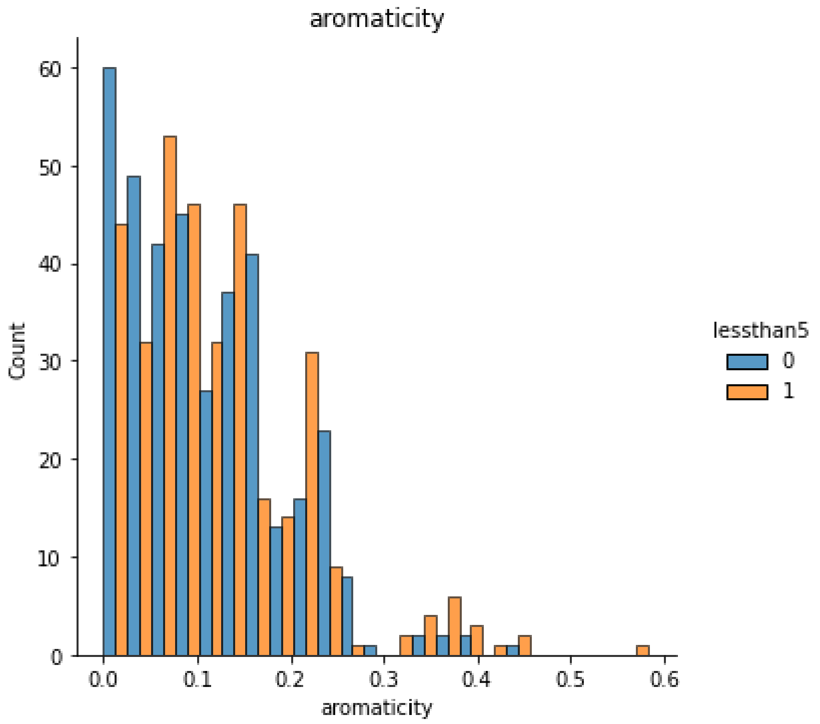

PCP_SA_BU
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
0.2
0.4
0.6
0.8
PCP_SA_BU
lessthan5
0
1

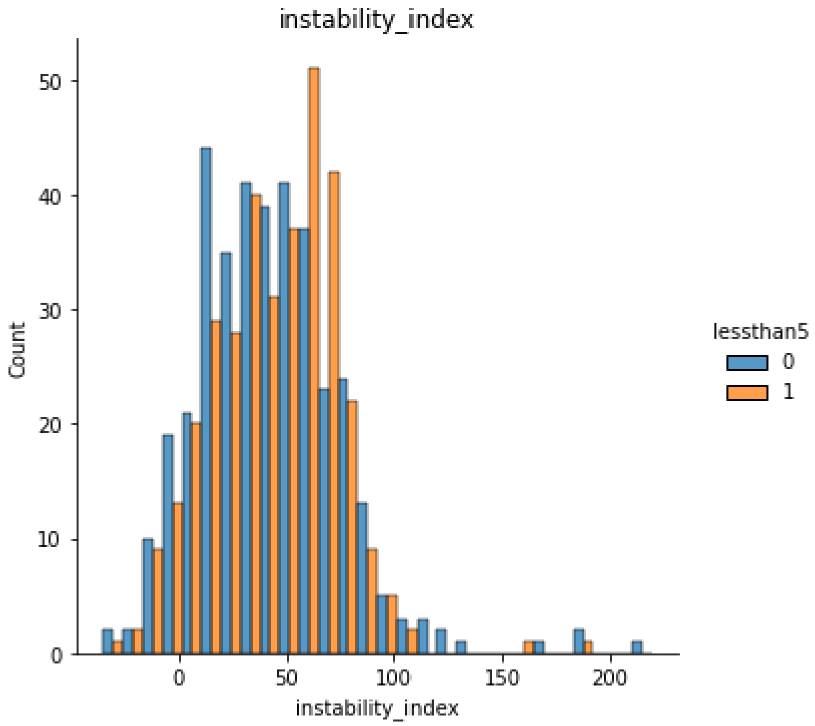

PCP_SA_EX
Proportion
PCP_SA_EX
lessthan5
0
1

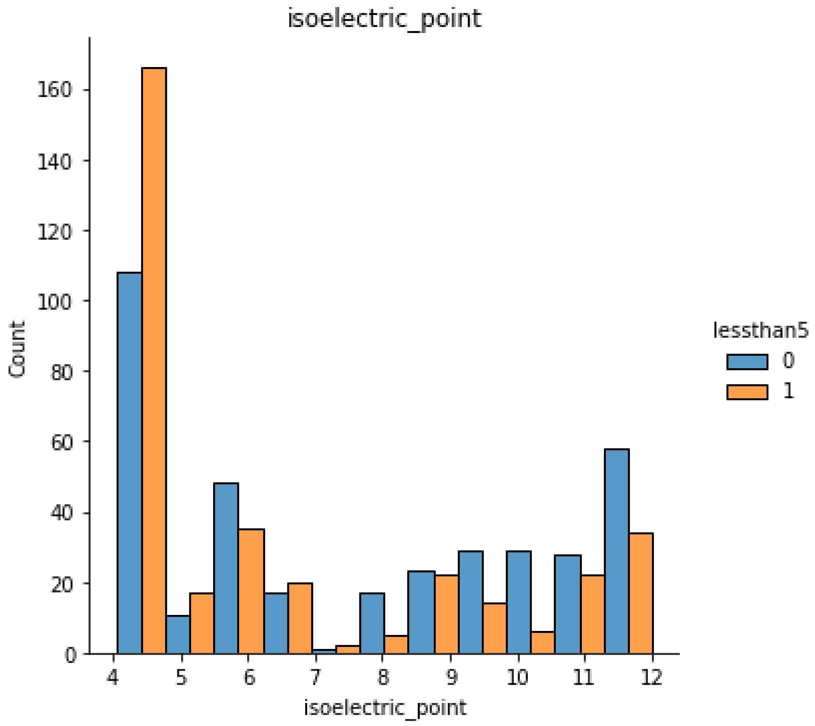

PCP_SA_IN
Proportion
PCP_SA_IN
lessthan5
0
1

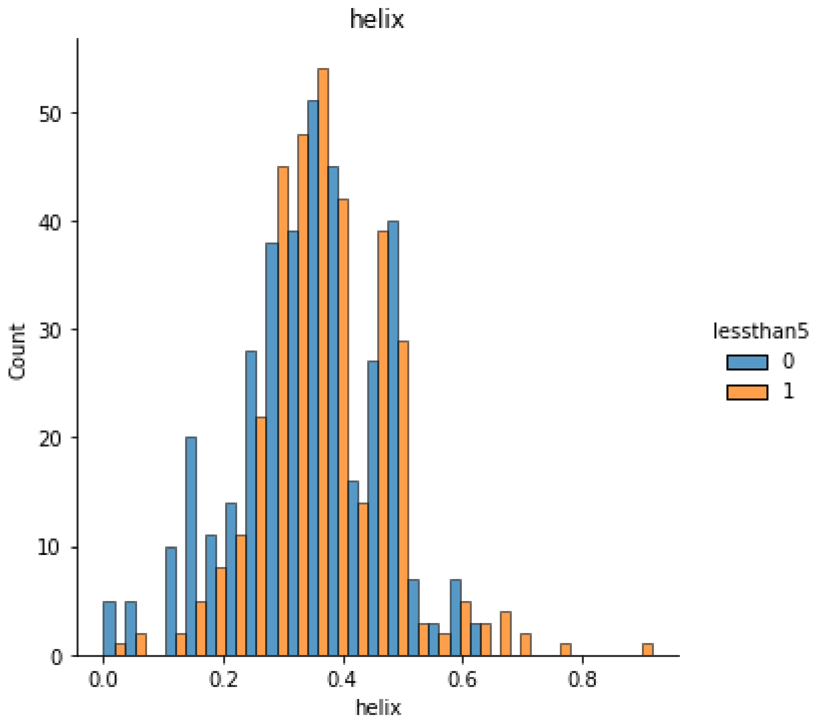
In [41]:
```
for feature in physicochem_prop.columns:
    sns.displot(x=physicochem_prop[feature], hue=df['lessthan5'], multip
le="dodge")
    plt.title(feature)
    plt.show()
```

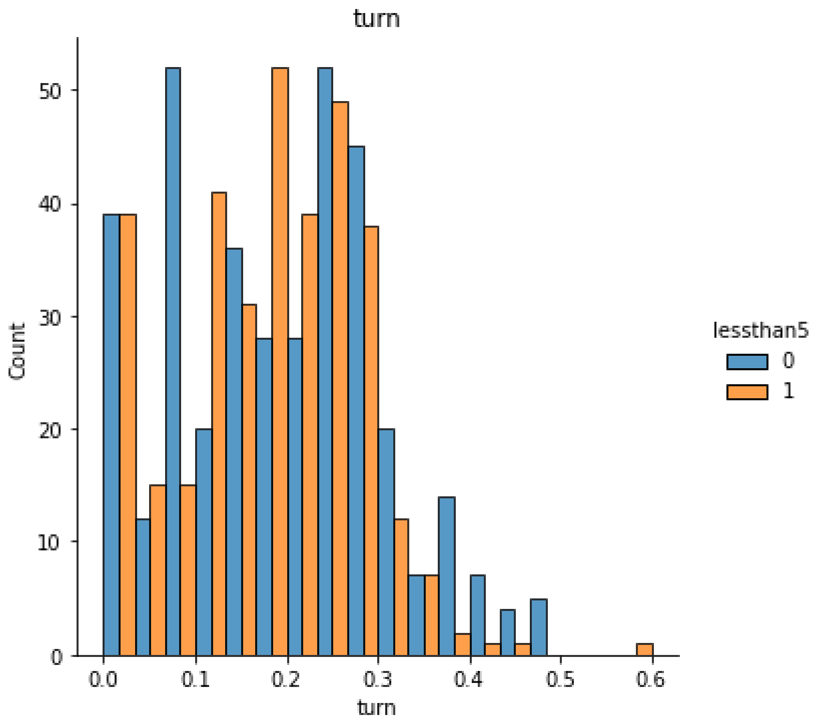

molecular_weight
Count
molecular_weight
lessthan5
0
1
0
20
40
60
80
1000
2000
3000
4000
5000
6000
7000

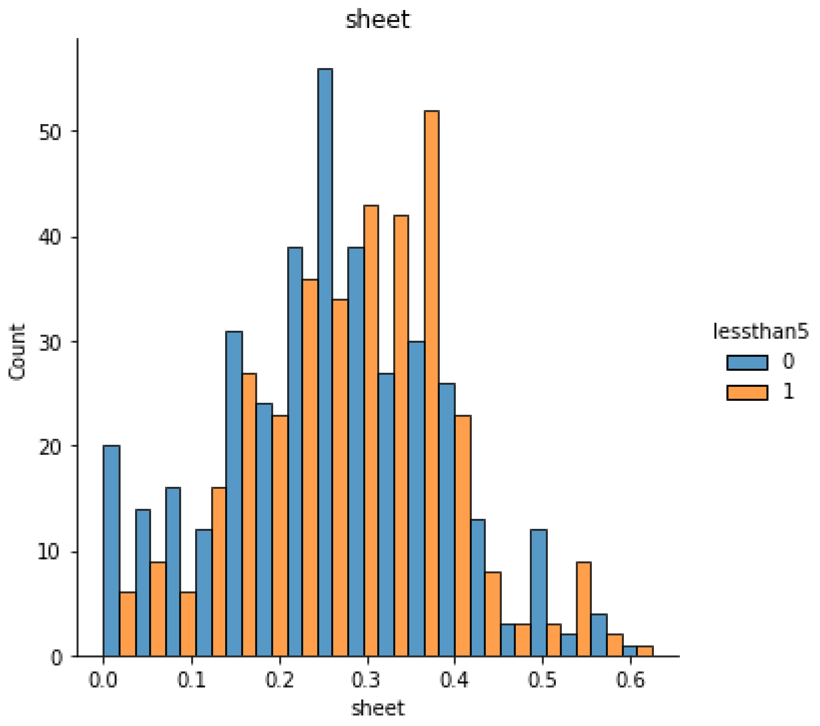

aromaticity
Count
aromaticity
lessthan5
0
1
0
10
20
30
40
50
60
0.0
0.1
0.2
0.3
0.4
0.5
0.6

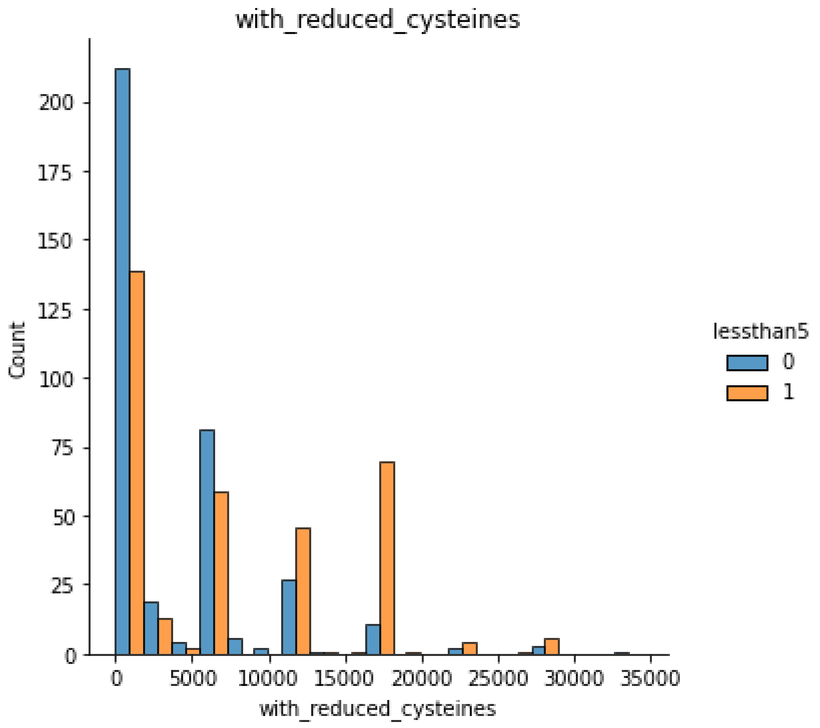

instability_index
Count
instability_index
lessthan5
0
1

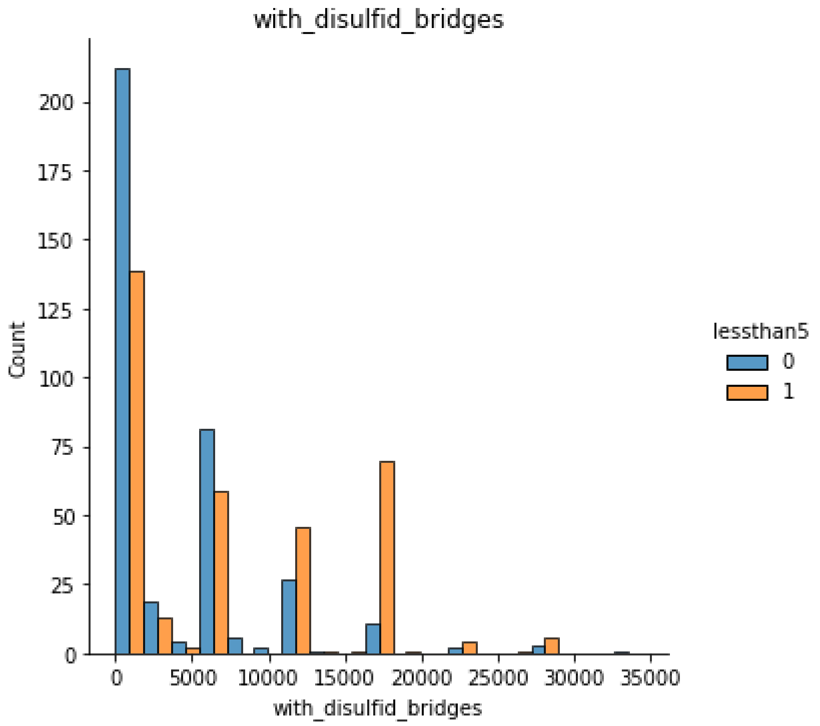

isoelectric_point
Count
isoelectric_point
lessthan5
0
1

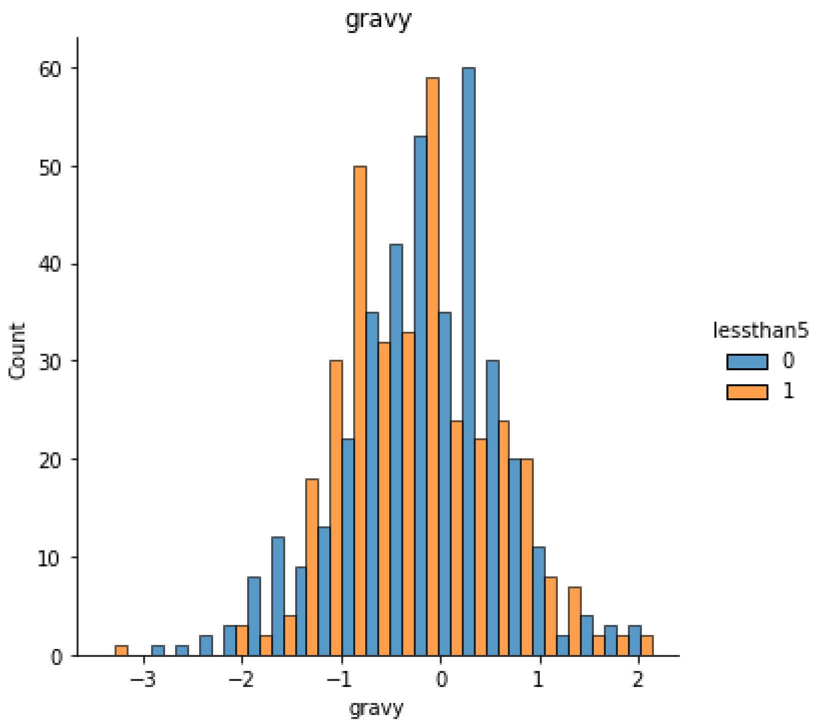

helix
Count
helix
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50

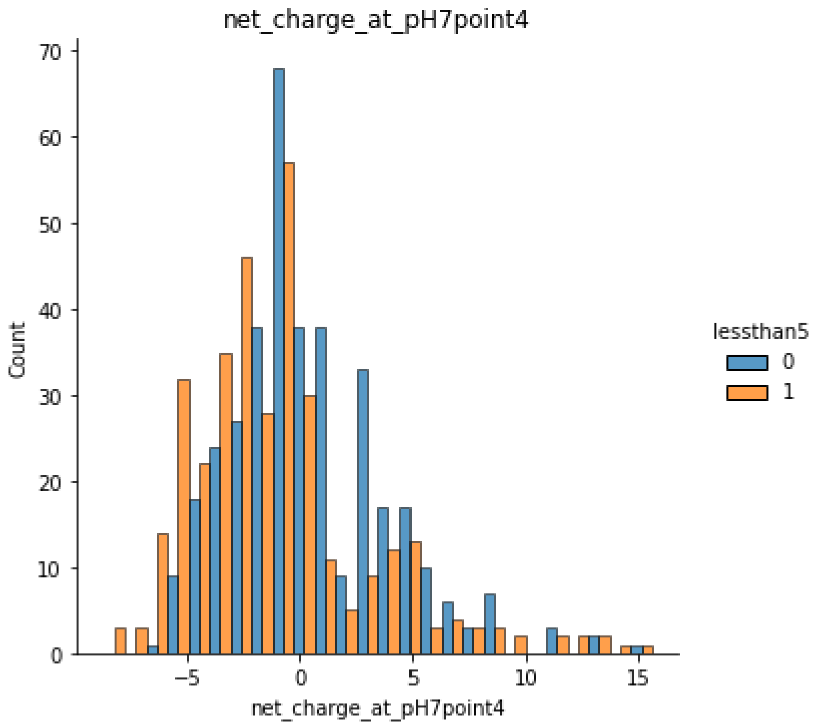

turn
Count
turn
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
50

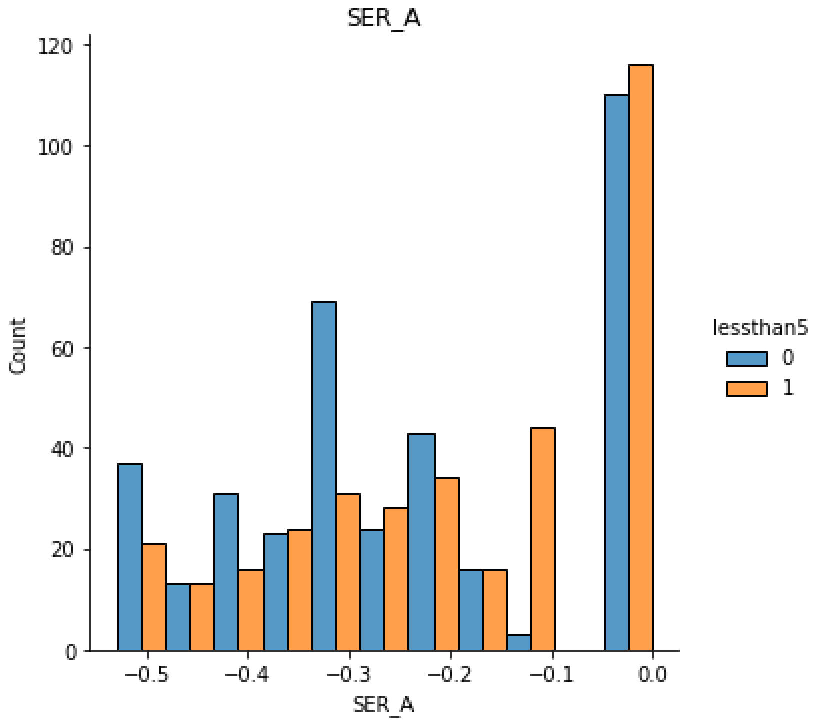

sheet
Count
sheet
lessthan5
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
50

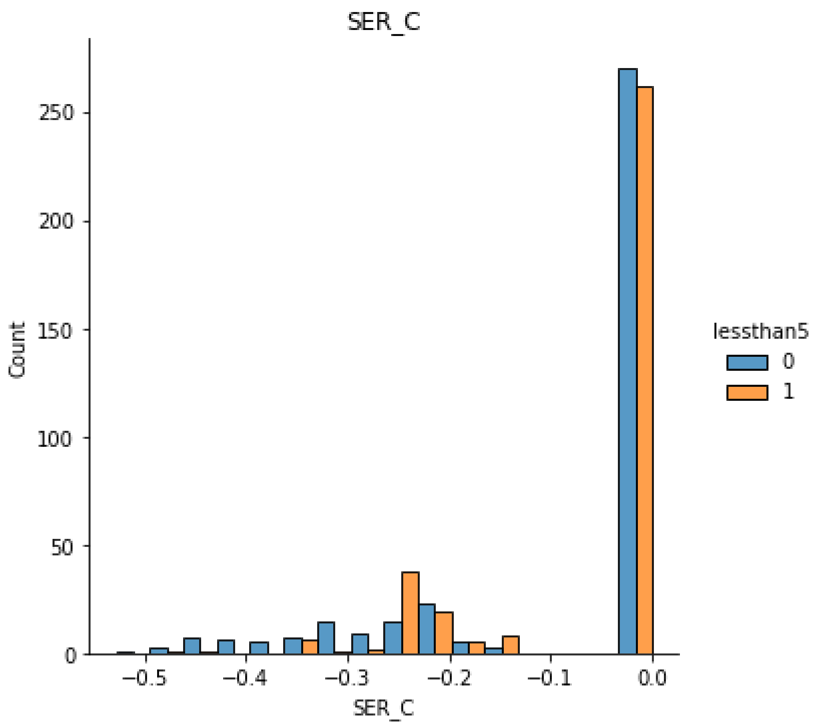

with_reduced_cysteines
Count
with_reduced_cysteines
lessthan5
0
1
0
5000
10000
15000
20000
25000
30000
35000
0
25
50
75
100
125
150
175
200

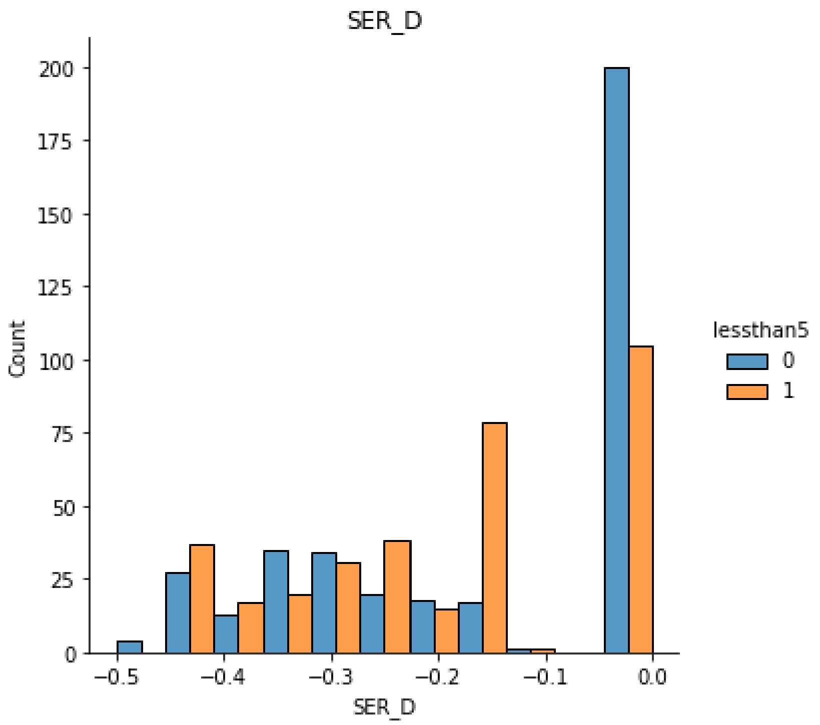

with_disulfid_bridges
Count
200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
30000
35000
with_disulfid_bridges
lessthan5
0
1

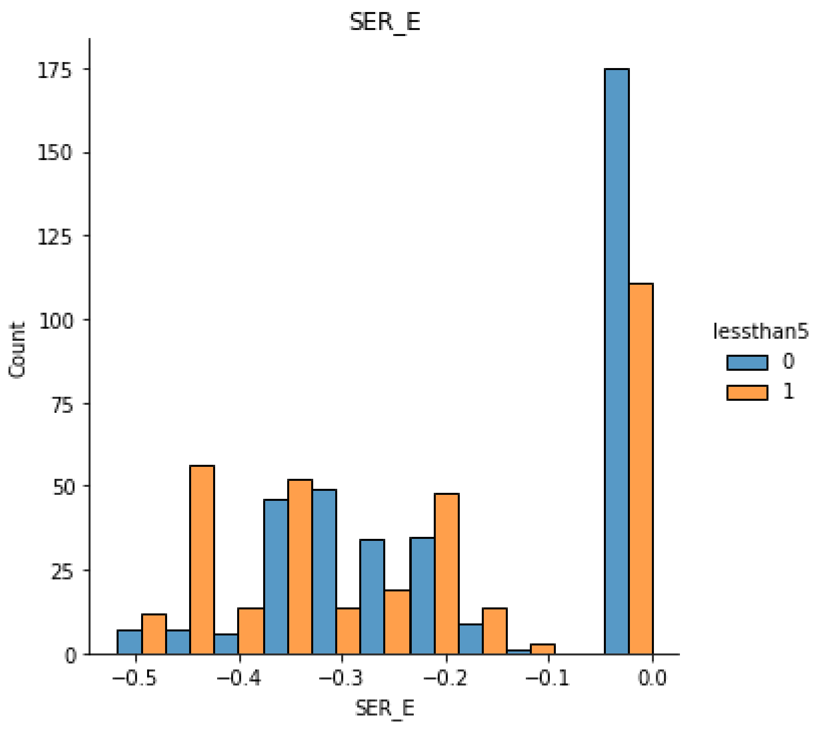

gravy
Count
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
gravy
lessthan5
0
1

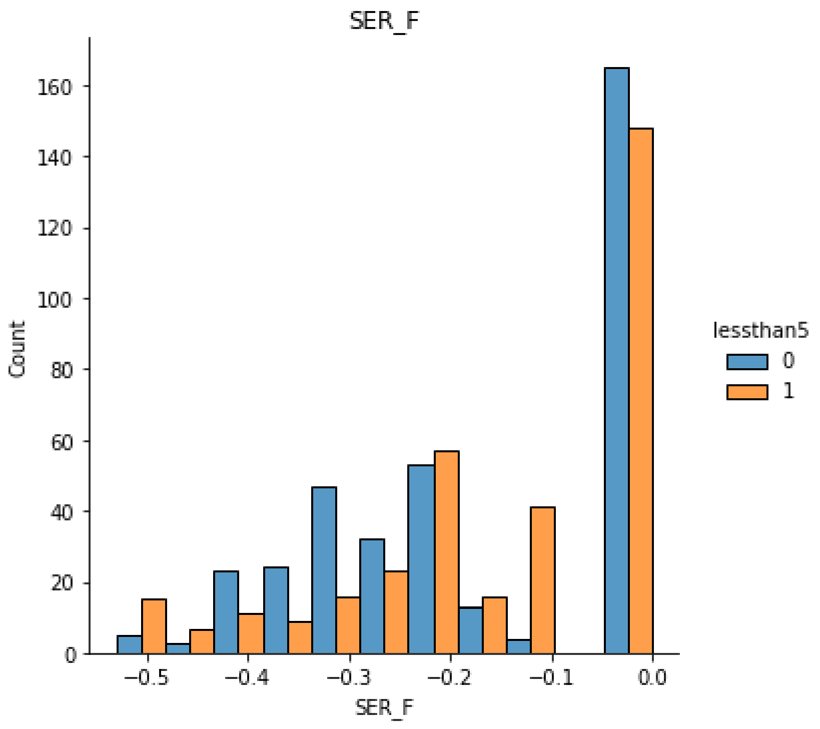

net_charge_at_pH7point4
Count
net_charge_at_pH7point4
lessthan5
0
1
70
60
50
40
30
20
10
0
−5
0
5
10
15

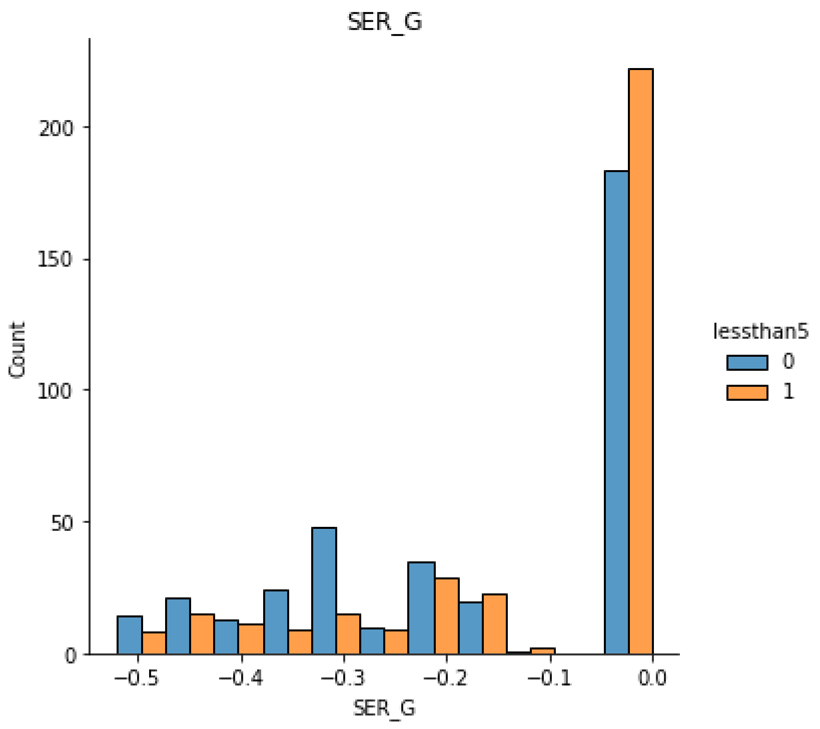

SER_A
Count
SER_A
lessthan5
0
1
120
100
80
60
40
20
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

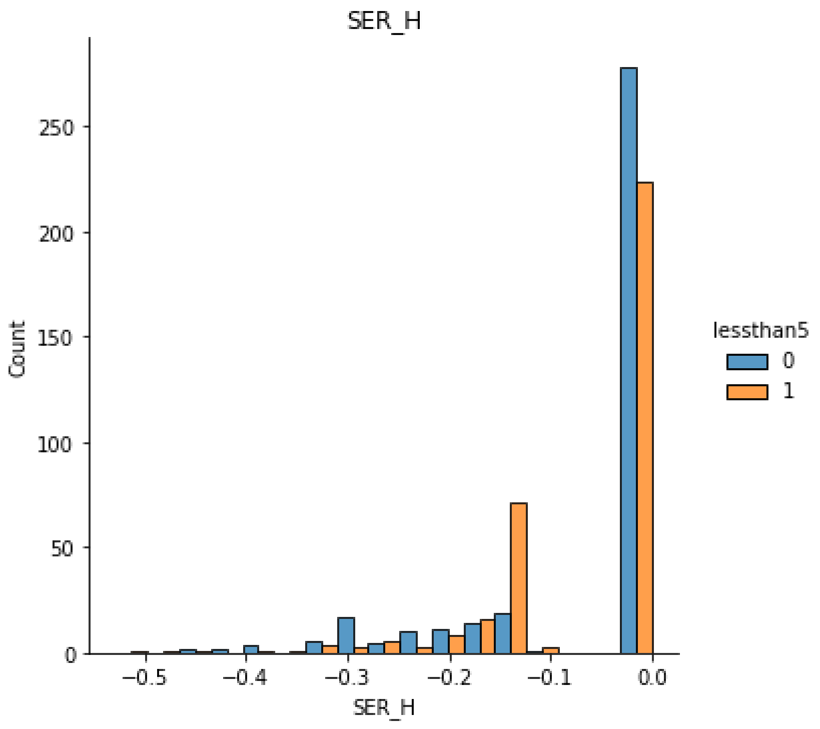

SER_C
Count
SER_C
lessthan5
0
1
0
50
100
150
200
250
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

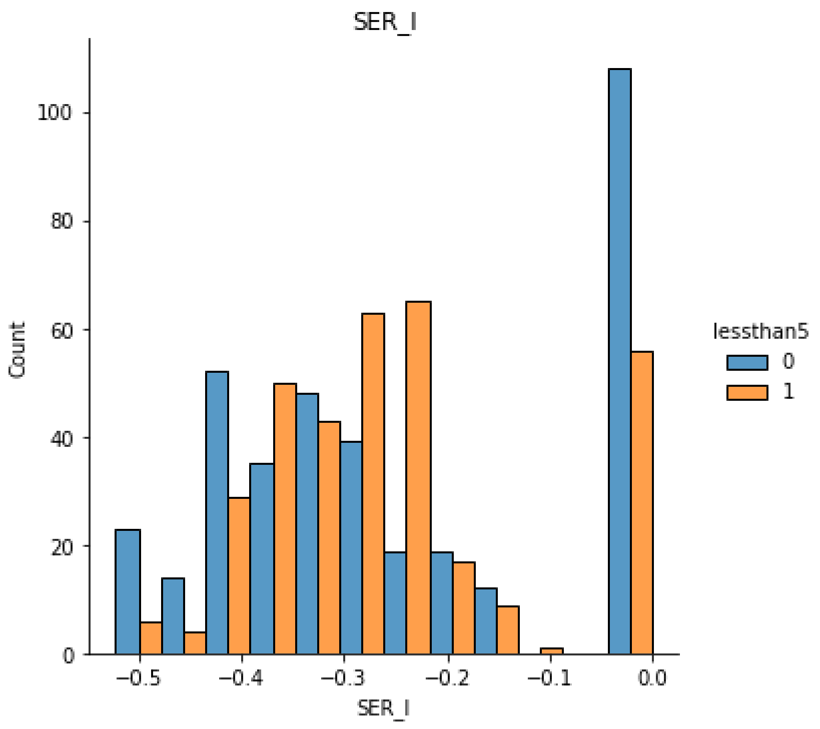

SER_D
Count
SER_D
lessthan5
0
1
0
25
50
75
100
125
150
175
200
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

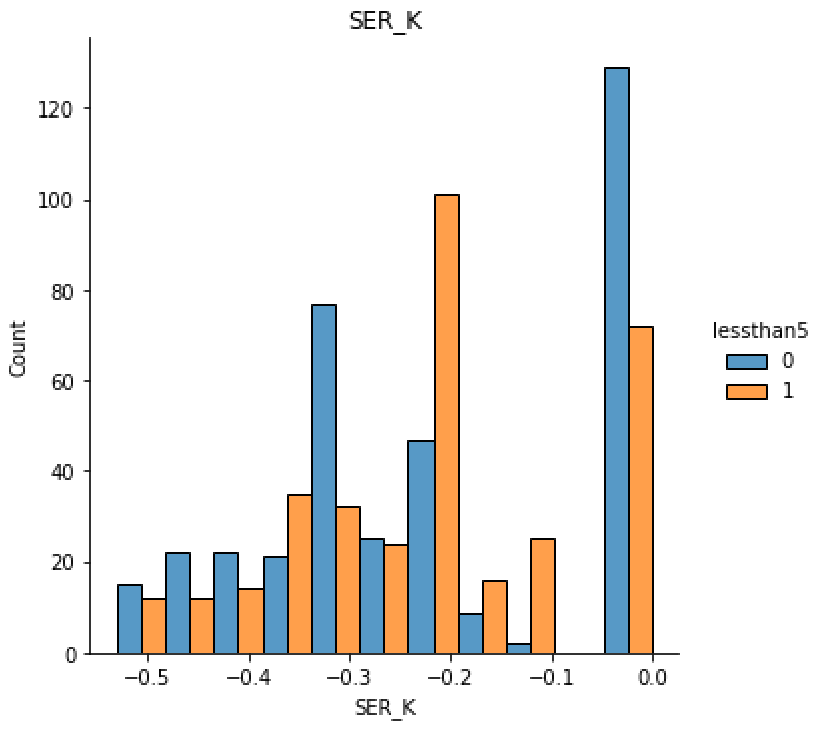

SER_E
Count
SER_E
lessthan5
0
1

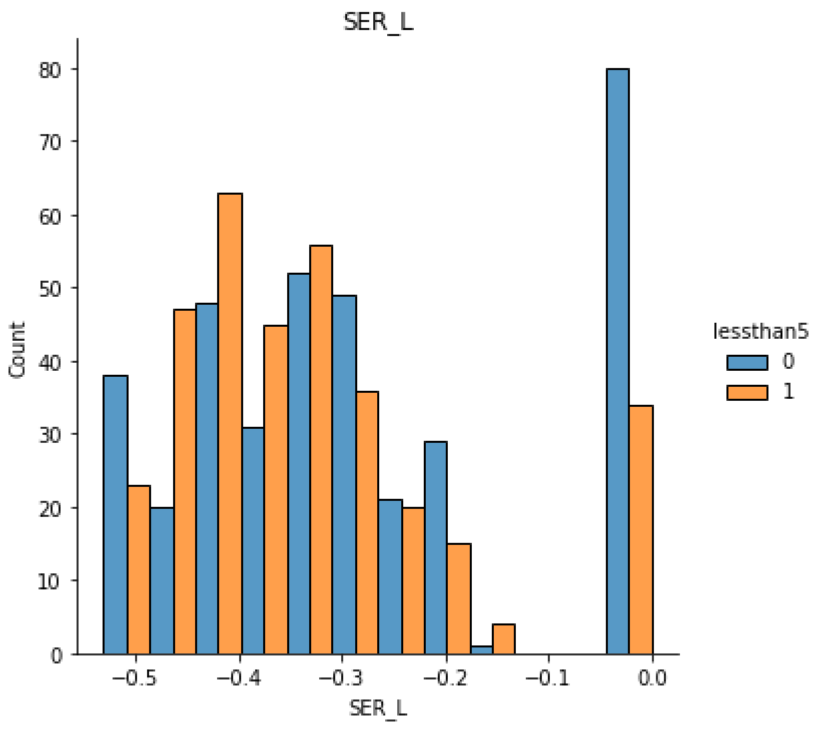

SER_F
Count
SER_F
lessthan5
0
1

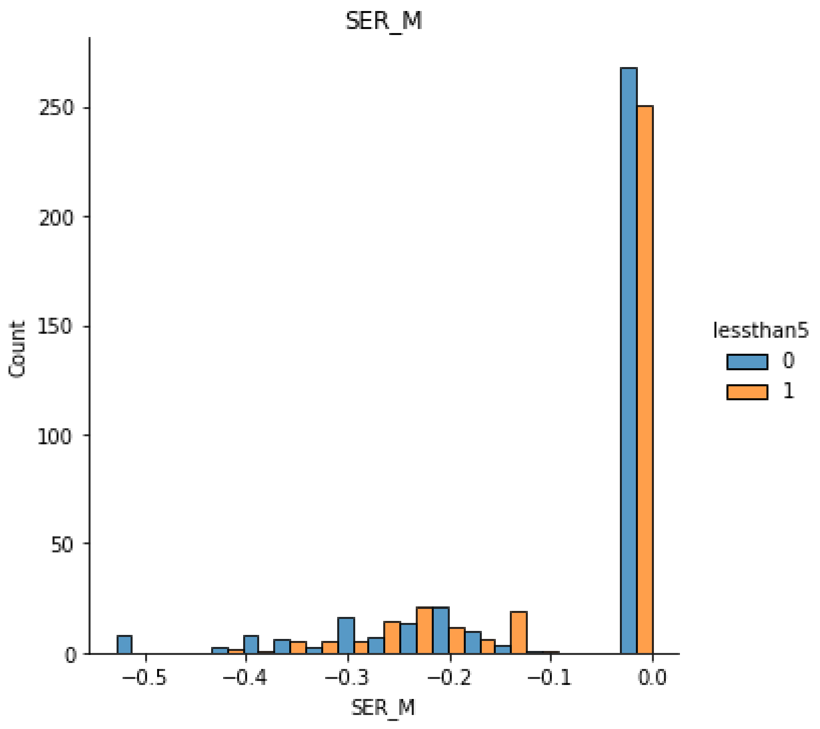

SER_G
Count
200
150
100
50
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_G
lessthan5
0
1

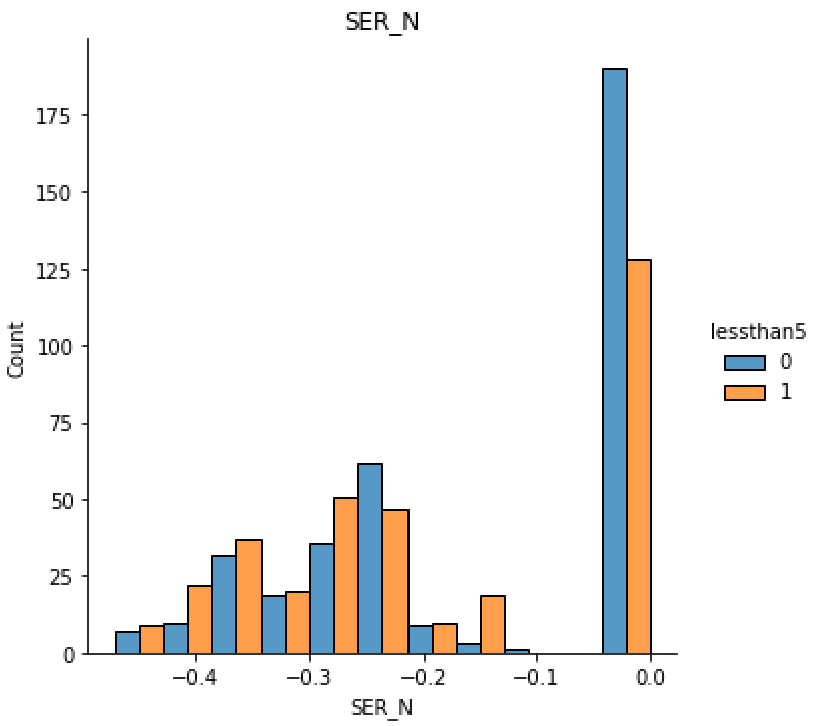

SER_H
Count
250
200
150
100
50
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_H
lessthan5
0
1

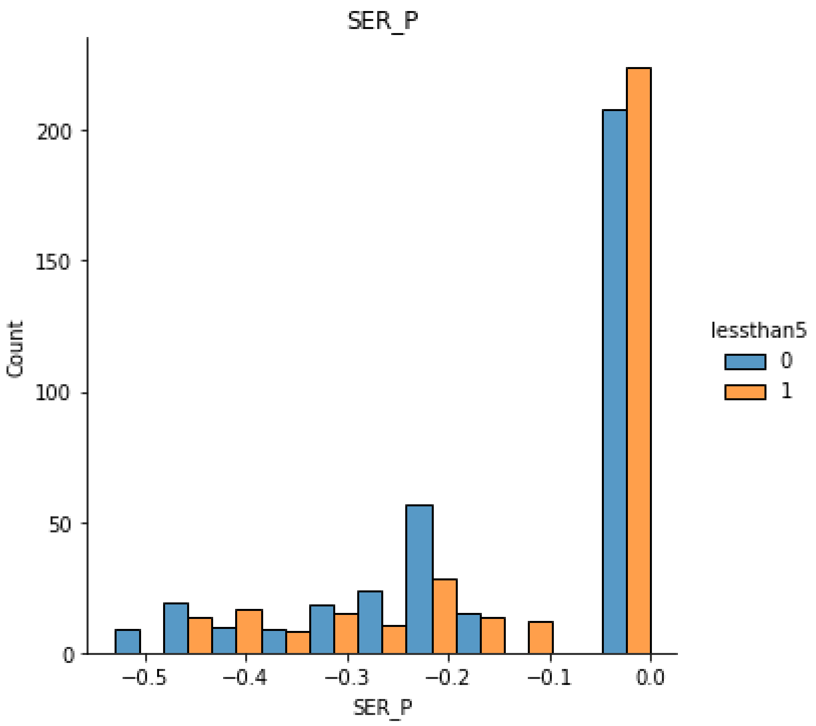

SER_I
Count
SER_I
lessthan5
0
1

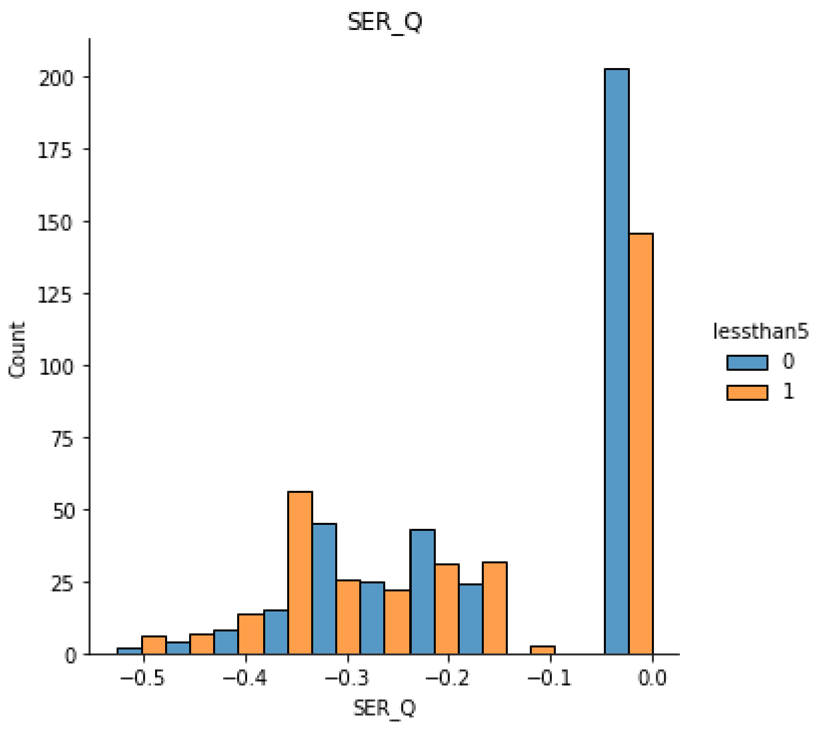

SER_K
Count
SER_K
lessthan5
0
1

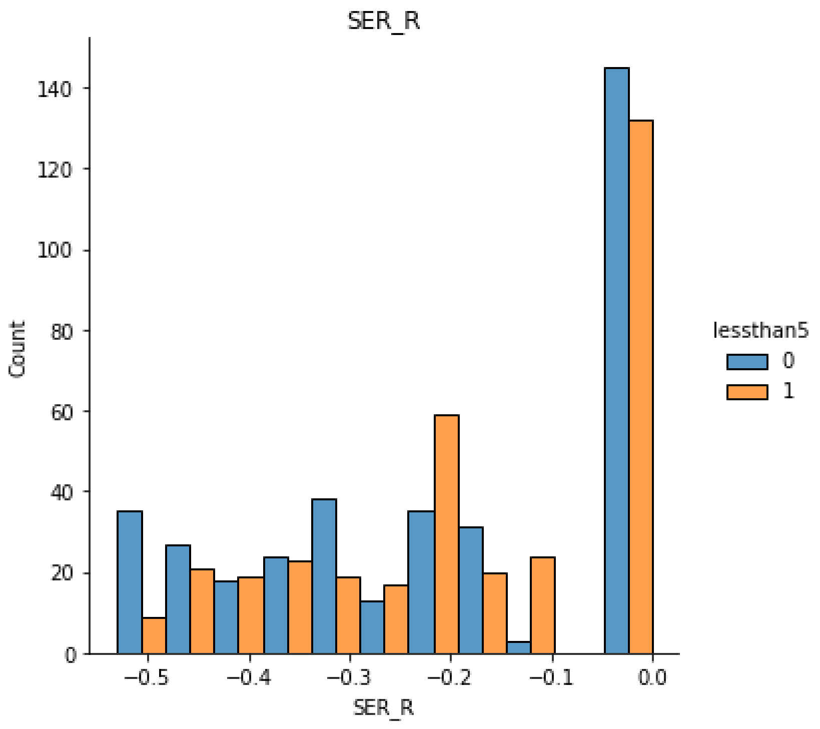

SER_L
Count
SER_L
lessthan5
0
1
80
70
60
50
40
30
20
10
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

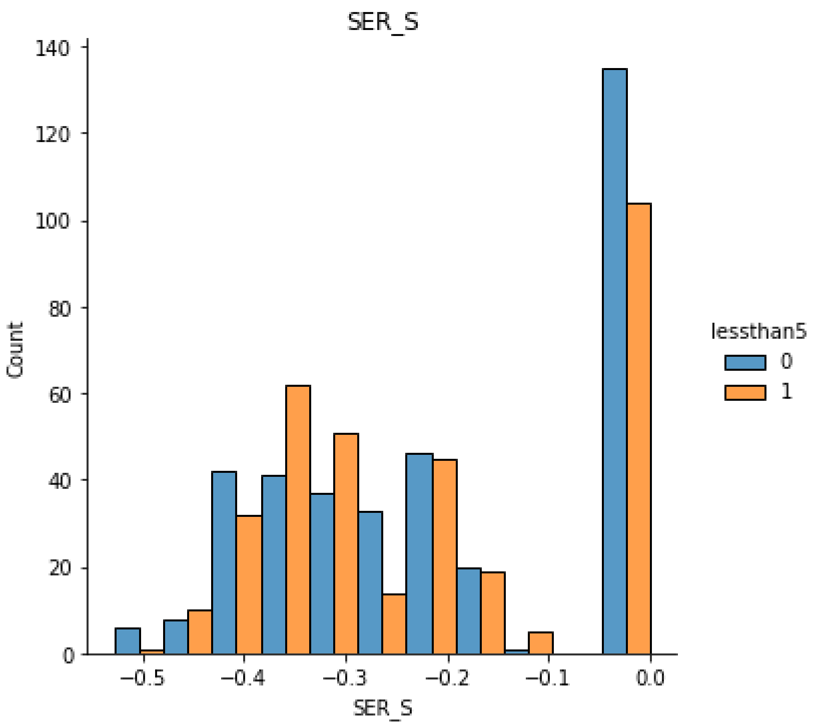

SER_M
Count
SER_M
lessthan5
0
1
250
200
150
100
50
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

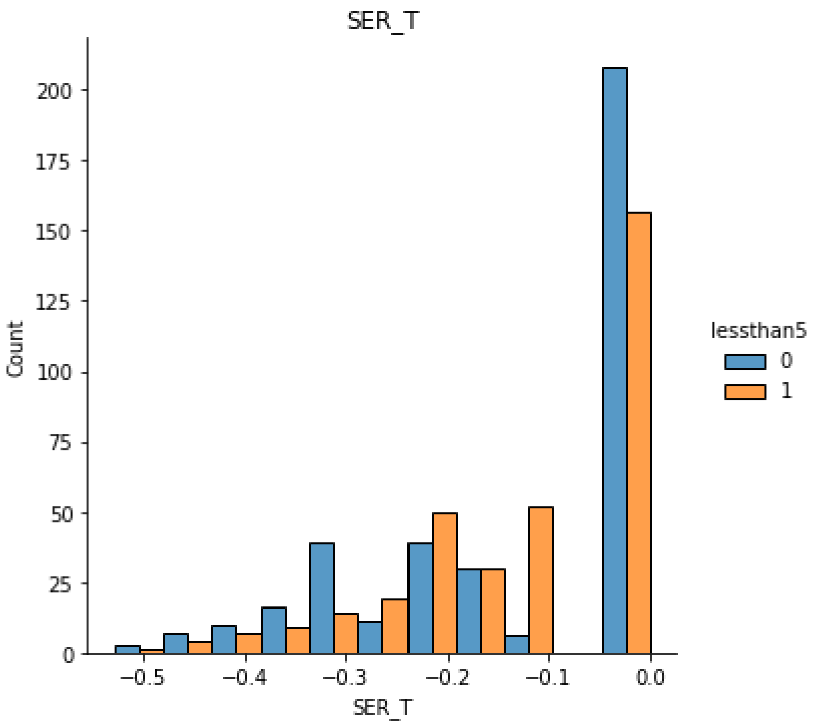

SER_N
Count
SER_N
lessthan5
0
1
175
150
125
100
75
50
25
0
−0.4
−0.3
−0.2
−0.1
0.0

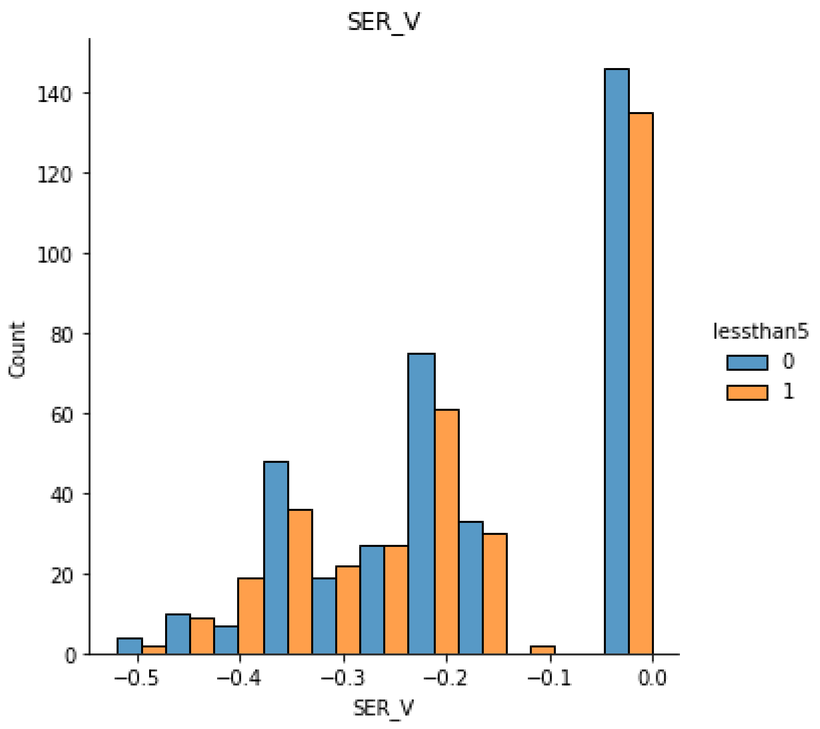

SER_P
Count
SER_P
lessthan5
0
1
200
150
100
50
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

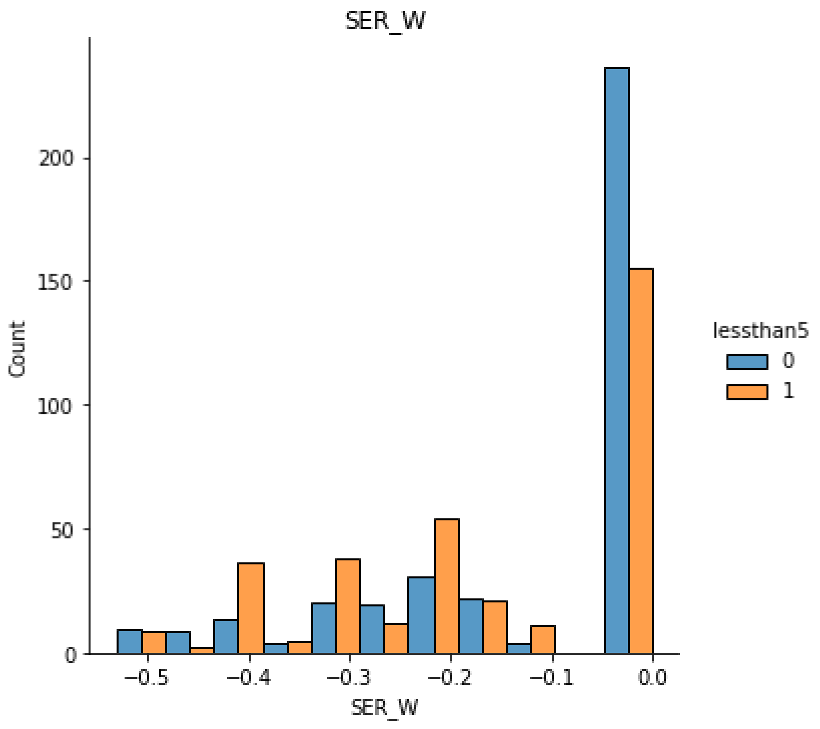

SER_Q
Count
200
175
150
125
100
75
50
25
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_Q
lessthan5
0
1

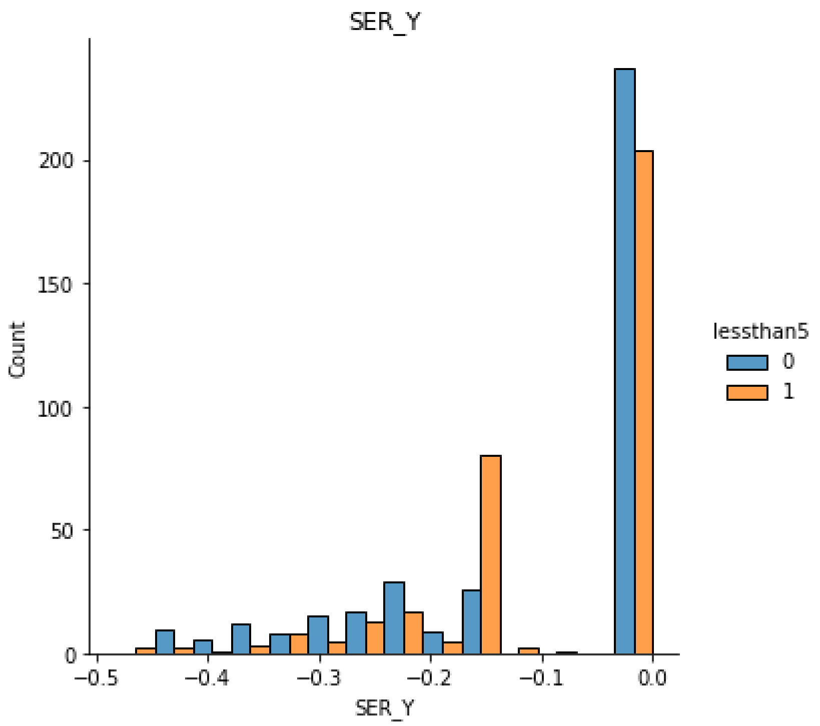

SER_R
Count
140
120
100
80
60
40
20
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
SER_R
lessthan5
0
1

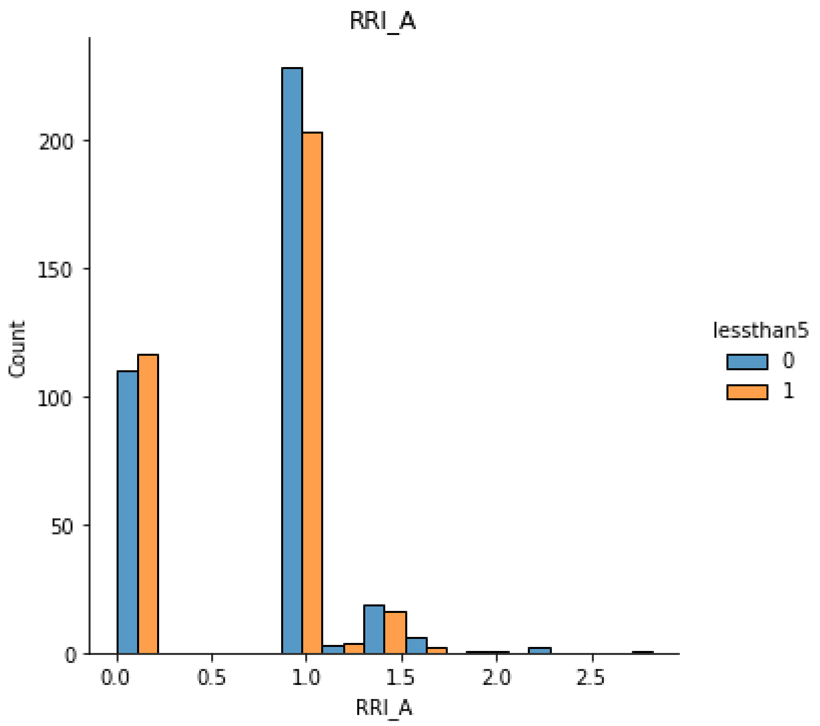

SER_S
Count
SER_S
lessthan5
0
1
140
120
100
80
60
40
20
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

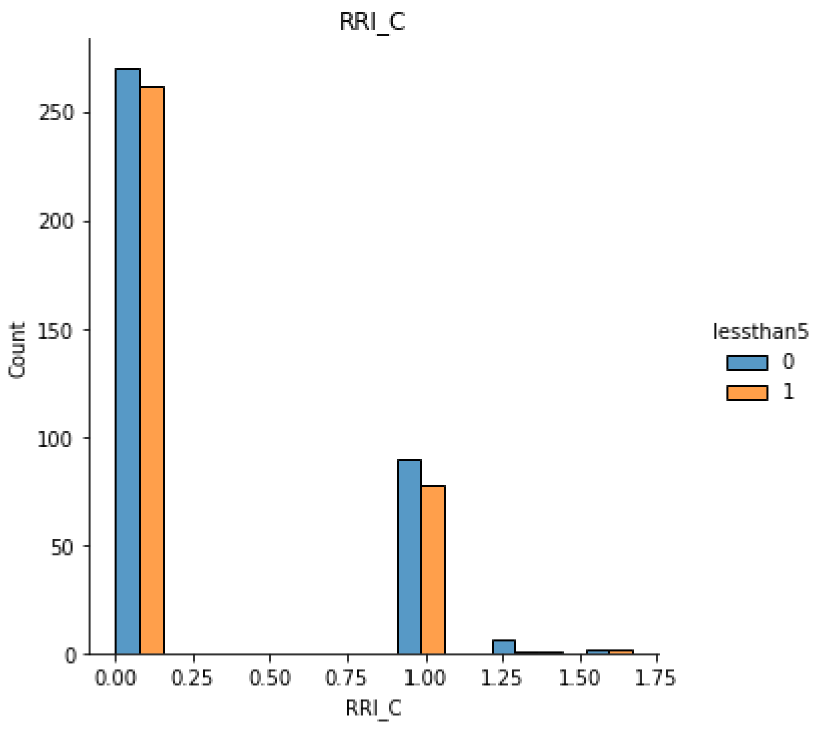

SER_T
Count
SER_T
lessthan5
0
1
200
175
150
125
100
75
50
25
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

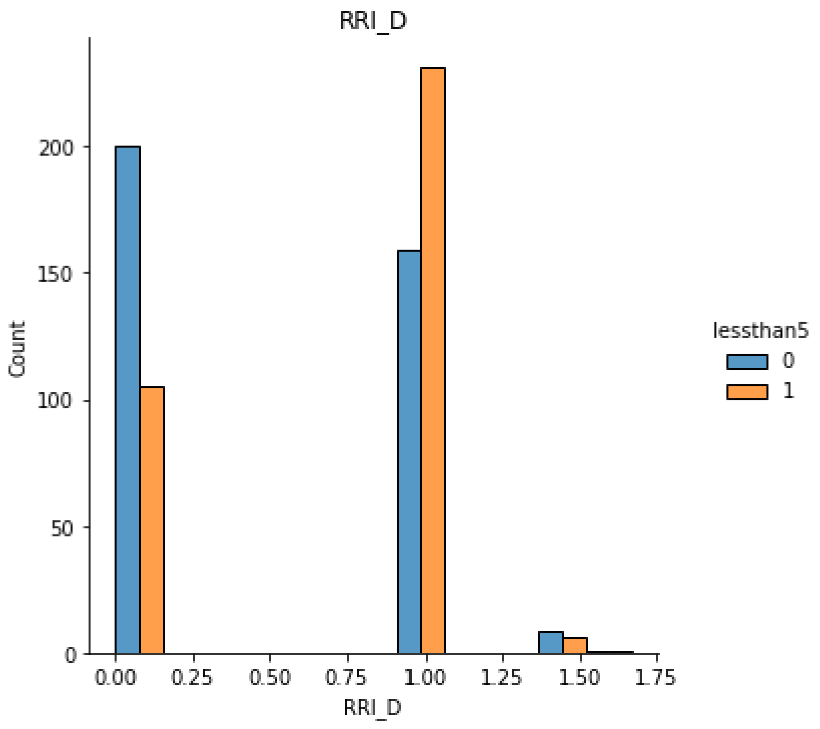

SER_V
Count
SER_V
lessthan5
0
1
140
120
100
80
60
40
20
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

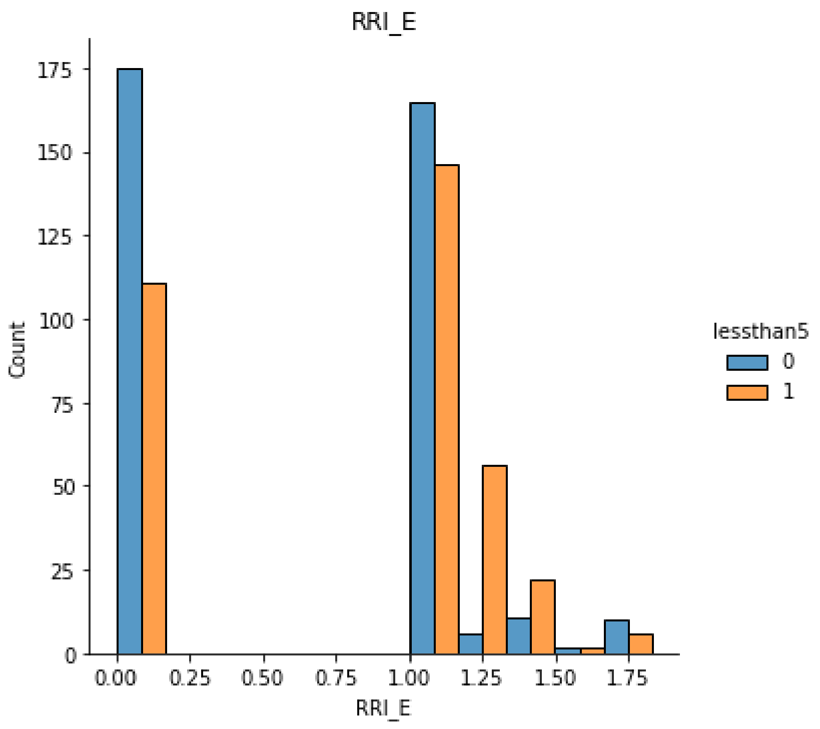

SER_W
Count
SER_W
lessthan5
0
1
200
150
100
50
0
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

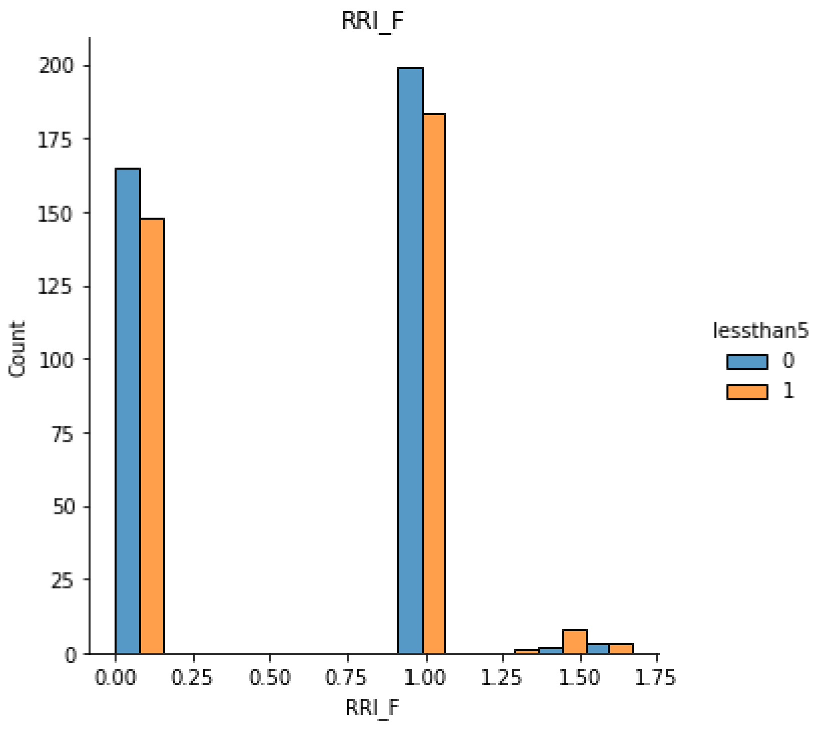

SER_Y
Count
SER_Y
lessthan5
0
1
0
50
100
150
200
−0.5
−0.4
−0.3
−0.2
−0.1
0.0

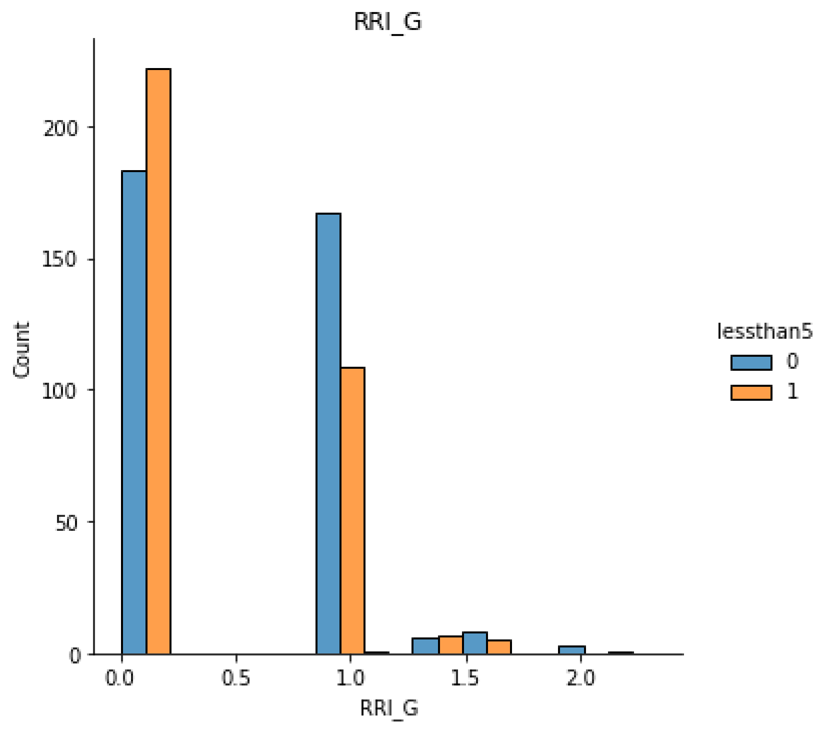

RRI_A
Count
RRI_A
lessthan5
0
1
0
50
100
150
200
0.0
0.5
1.0
1.5
2.0
2.5

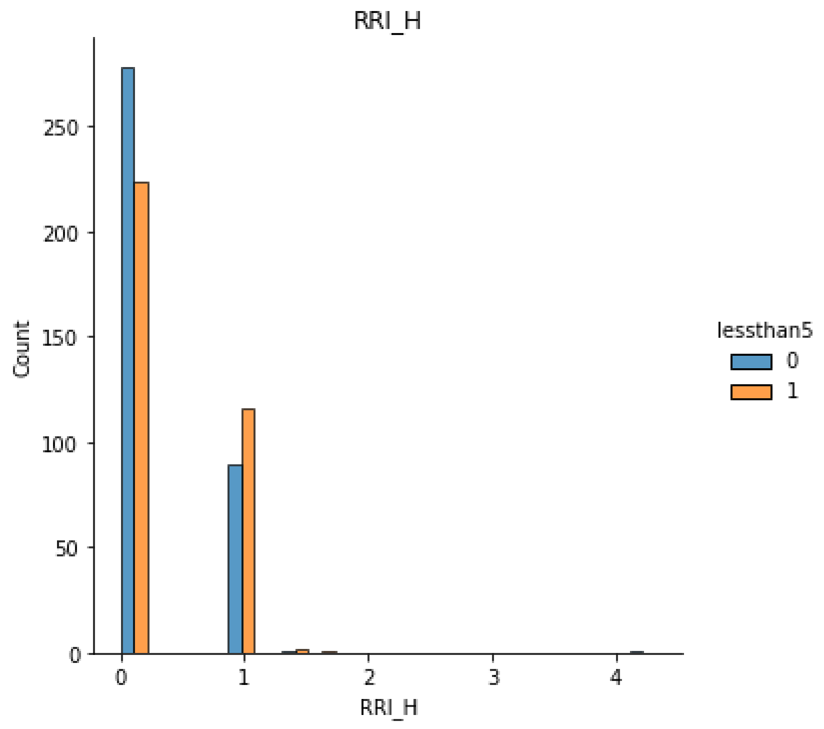

RRI_C
Count
250
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
RRI_C
lessthan5
0
1

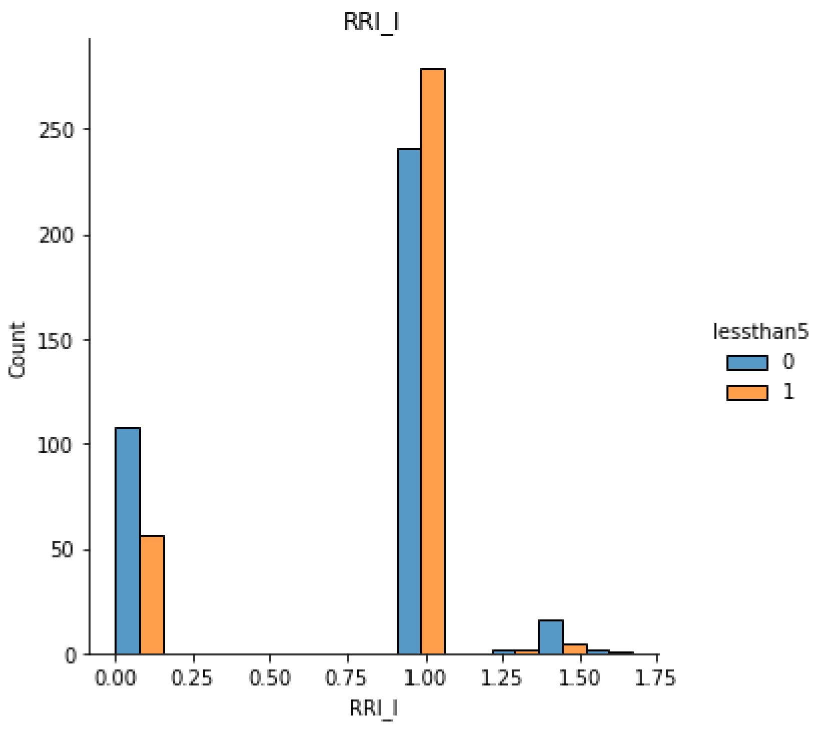

RRI_D
Count
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
RRI_D
lessthan5
0
1

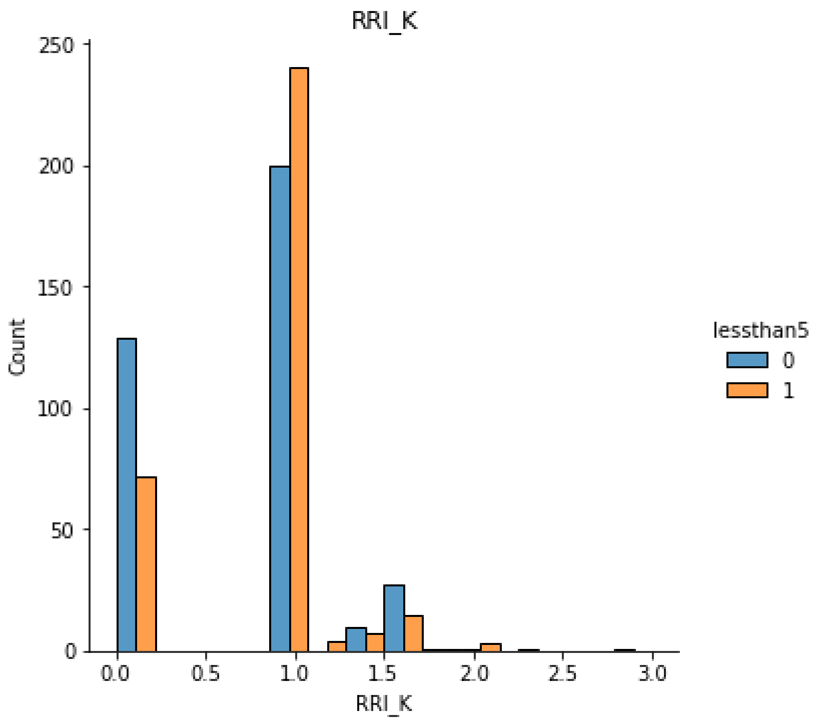

RRI_E
Count
RRI_E
lessthan5
0
1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
25
50
75
100
125
150
175

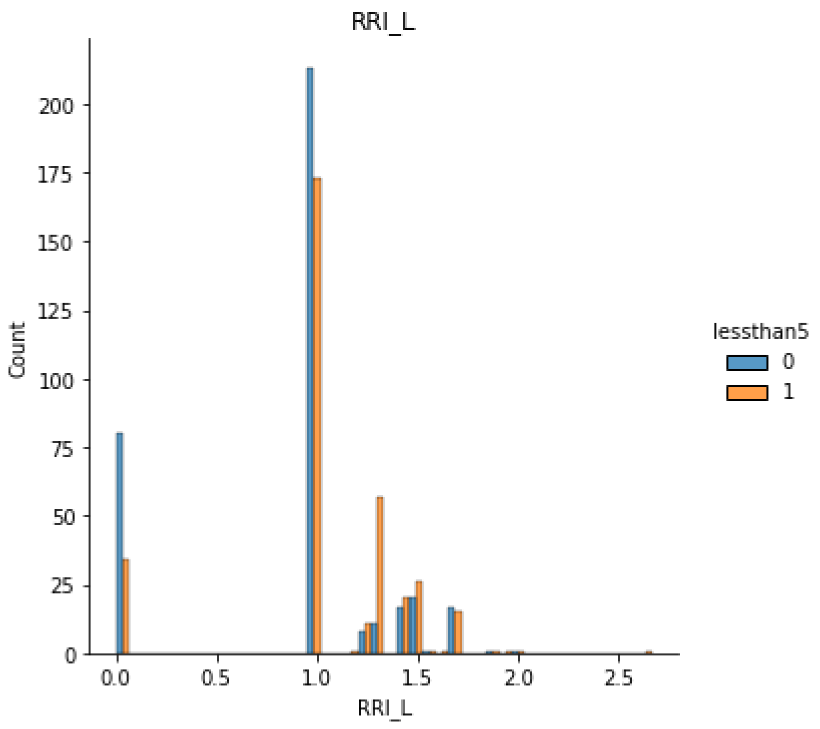

RRI_F
Count
RRI_F
lessthan5
0
1
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
25
50
75
100
125
150
175
200

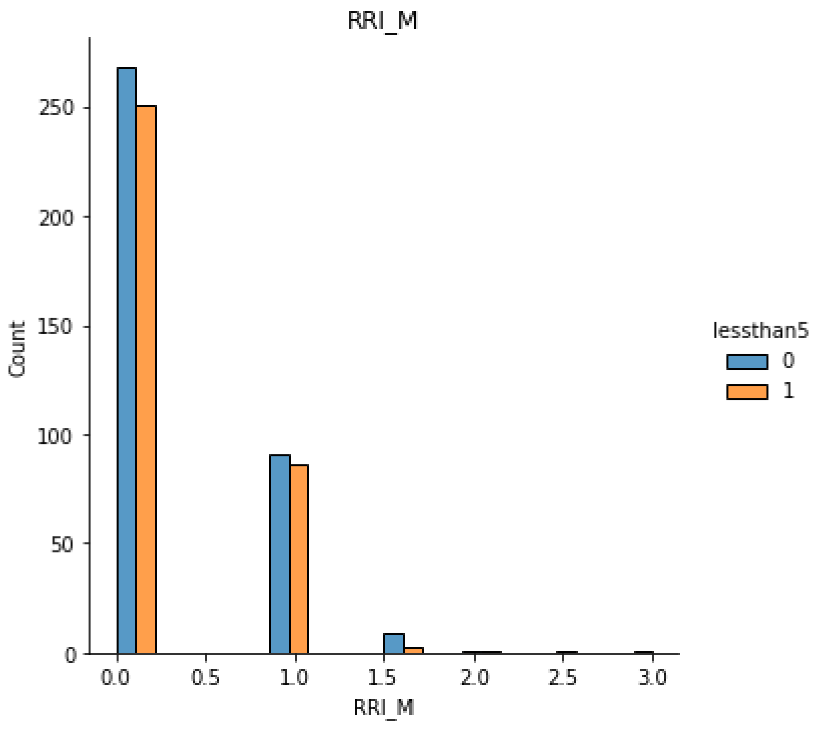

RRI_G
Count
RRI_G
lessthan5
0
1

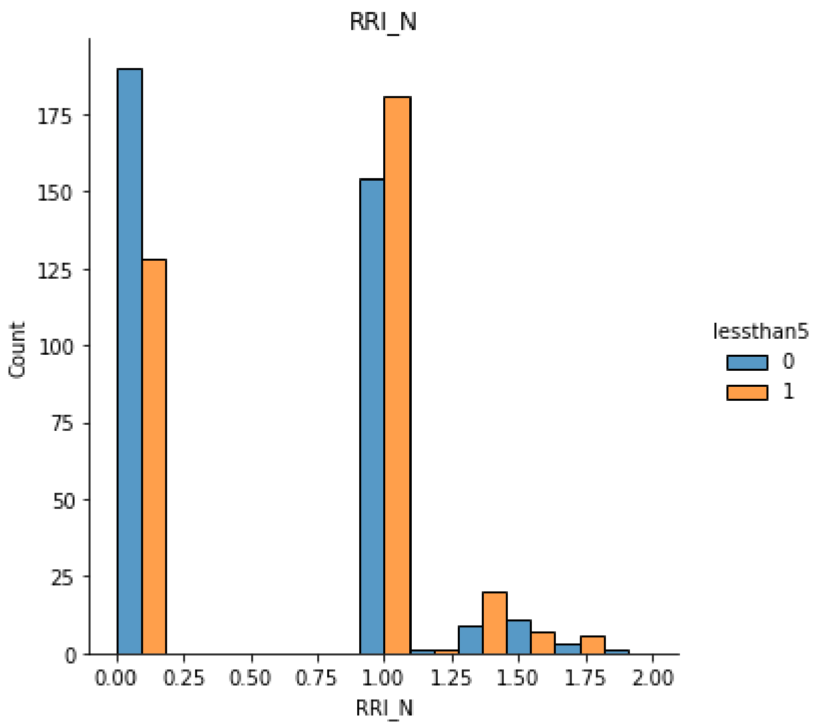

RRI_H
Count
RRI_H
lessthan5
0
1

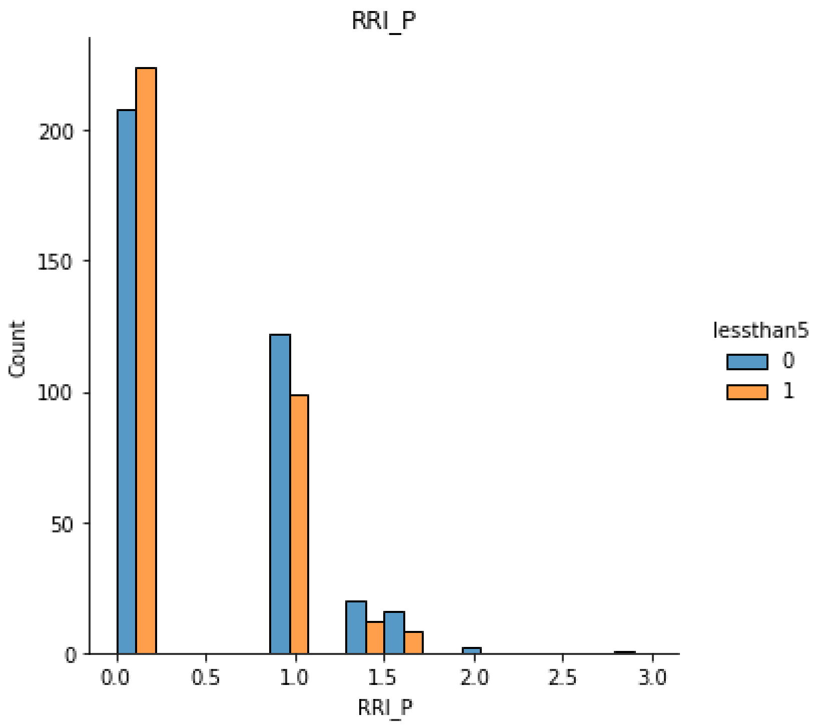

RRI_I
Count
RRI_I
250
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
lessthan5
0
1

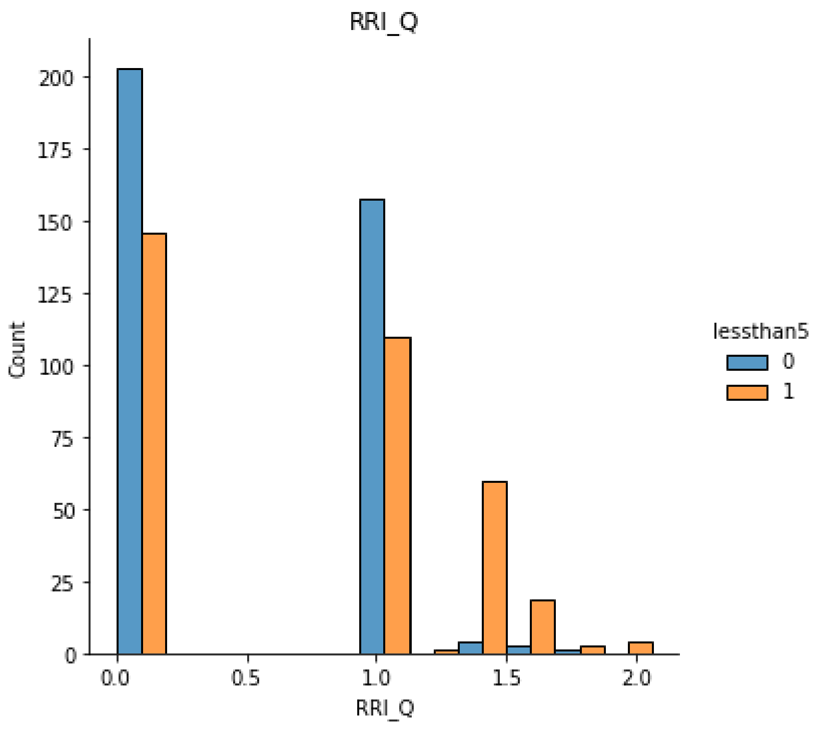

RRI_K
Count
RRI_K
250
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
lessthan5
0
1

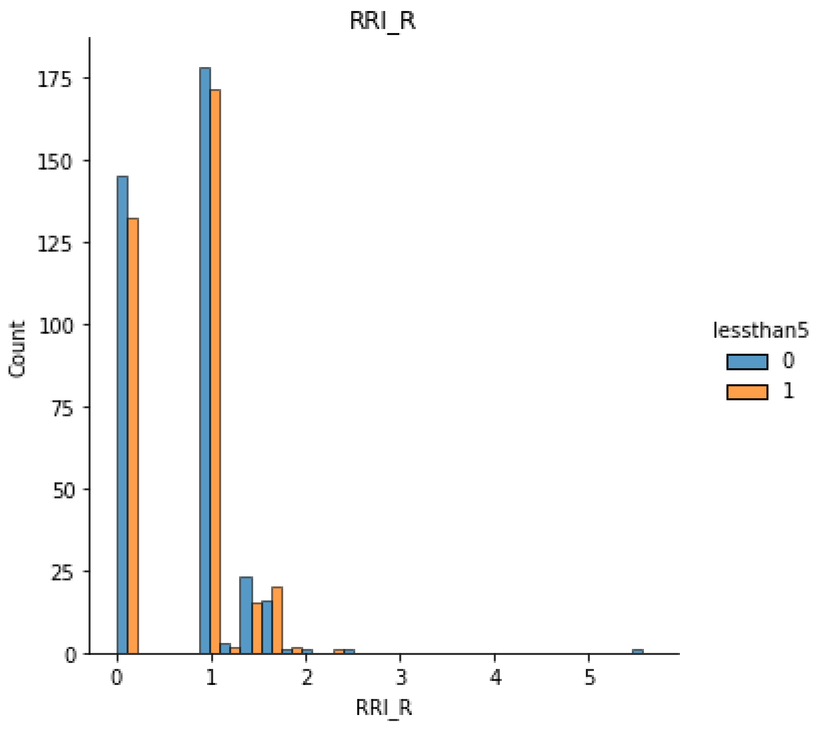

RRI_L
Count
RRI_L
lessthan5
0
1
0
25
50
75
100
125
150
175
200
0.0
0.5
1.0
1.5
2.0
2.5

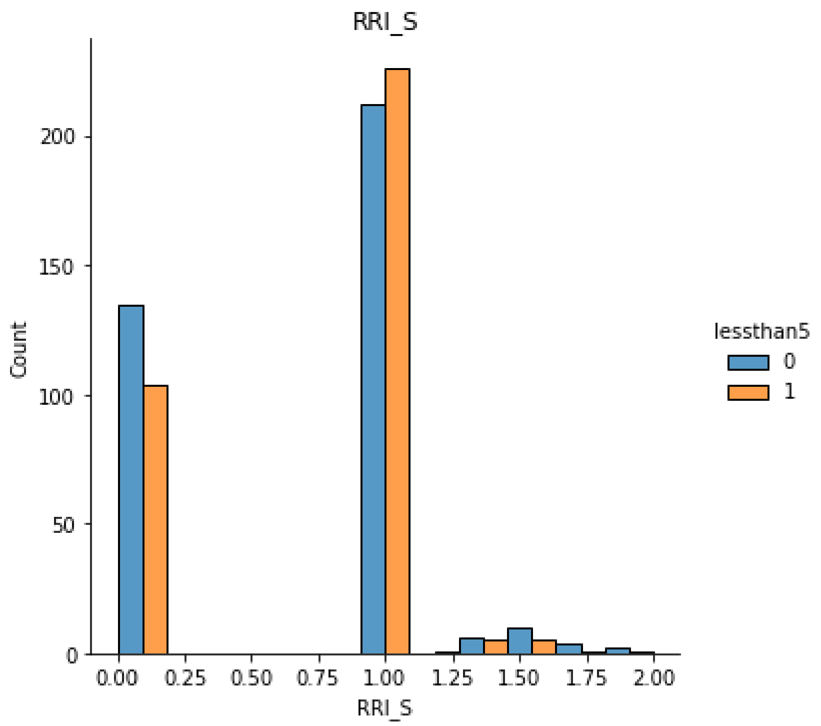

RRI_M
Count
RRI_M
lessthan5
0
1
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
2.5
3.0

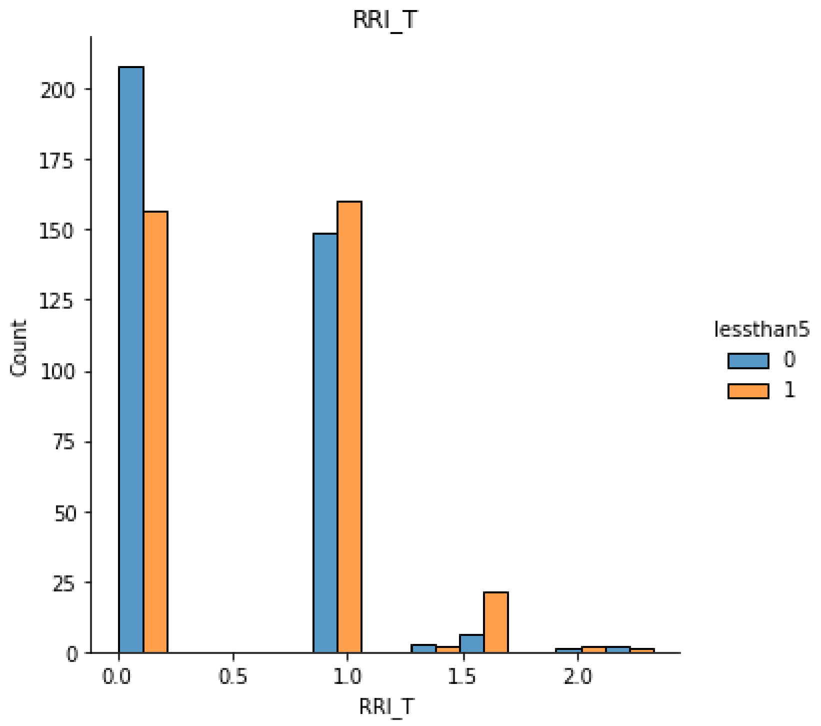

RRI_N
Count
175
150
125
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
RRI_N
lessthan5
0
1

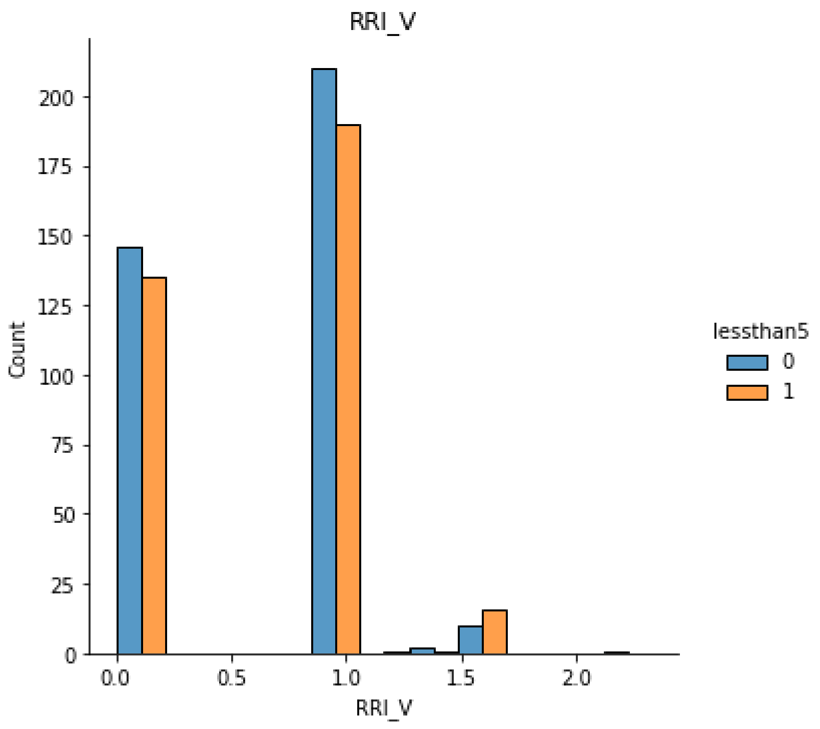

RRI_P
Count
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RRI_P
lessthan5
0
1

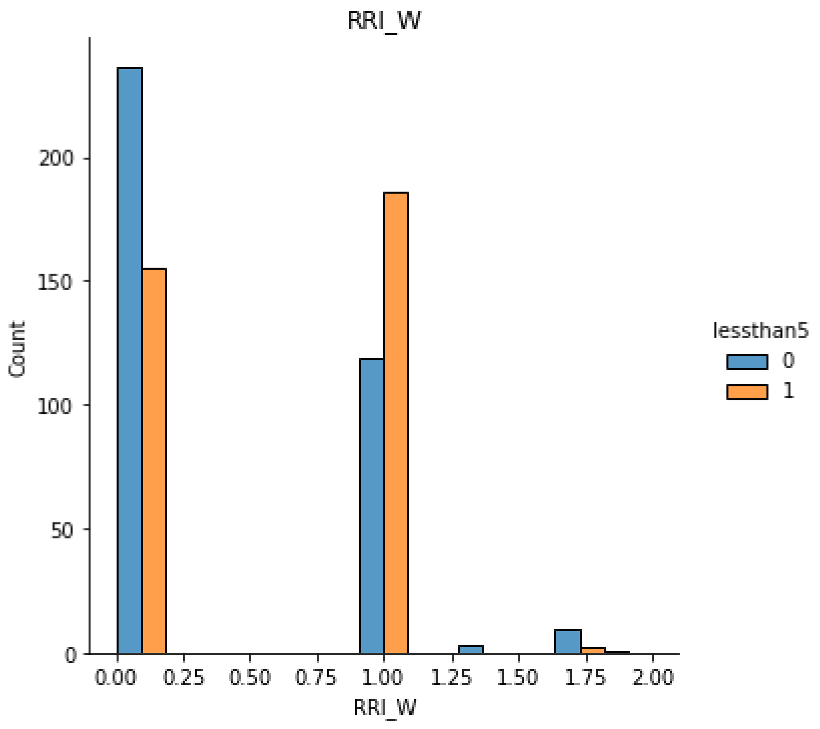

RRI_Q
Count
RRI_Q
lessthan5
0
1
0
25
50
75
100
125
150
175
200
0.0
0.5
1.0
1.5
2.0

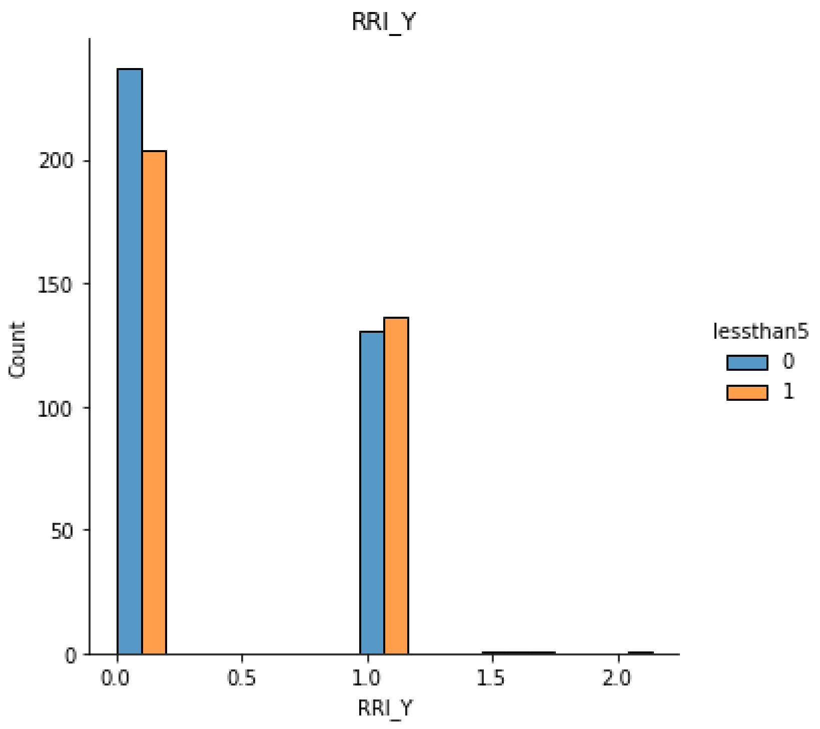

RRI_R
Count
RRI_R
lessthan5
0
1
0
25
50
75
100
125
150
175
0
1
2
3
4
5

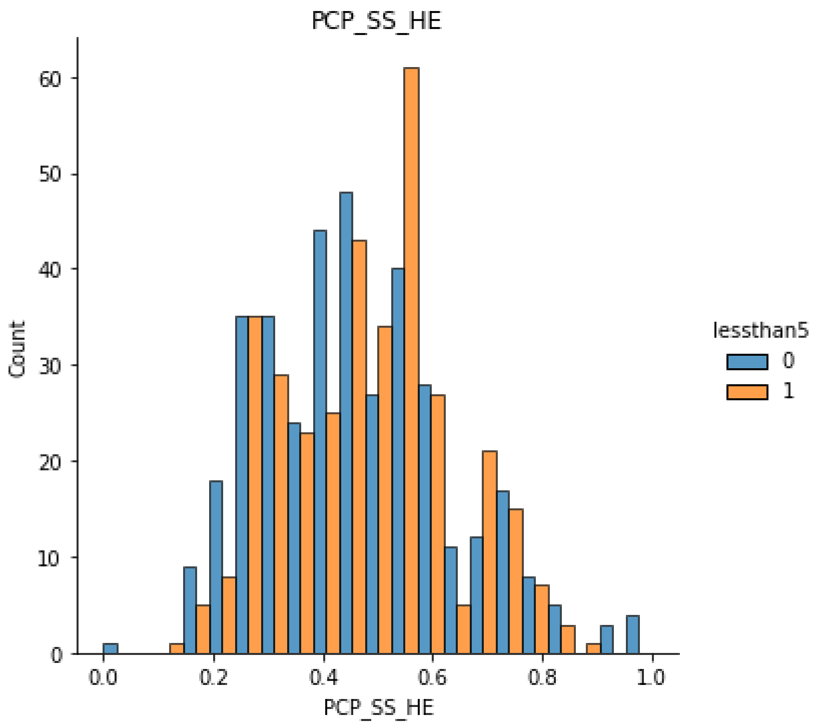

RRI_S
Count
RRI_S
lessthan5
0
1
0
50
100
150
200
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

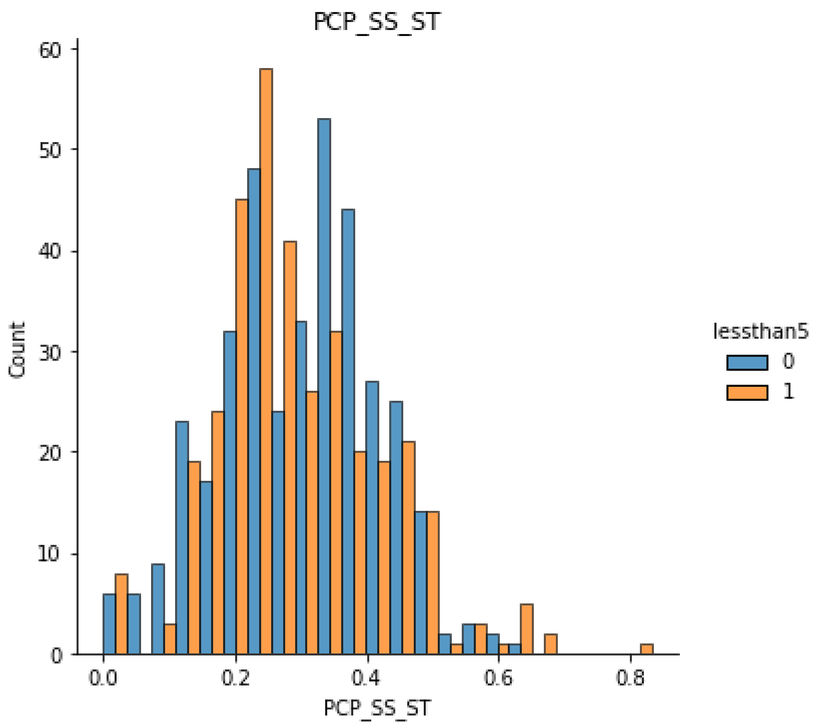

RRI_T
Count
RRI_T
lessthan5
0
1
0
25
50
75
100
125
150
175
200
0.0
0.5
1.0
1.5
2.0

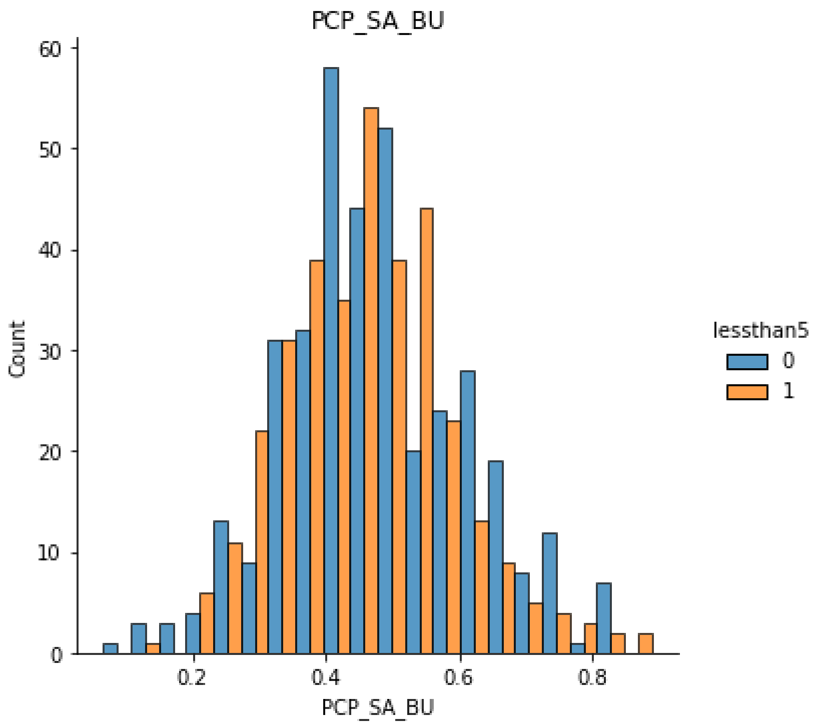

RRI_V
Count
RRI_V
lessthan5
0
1

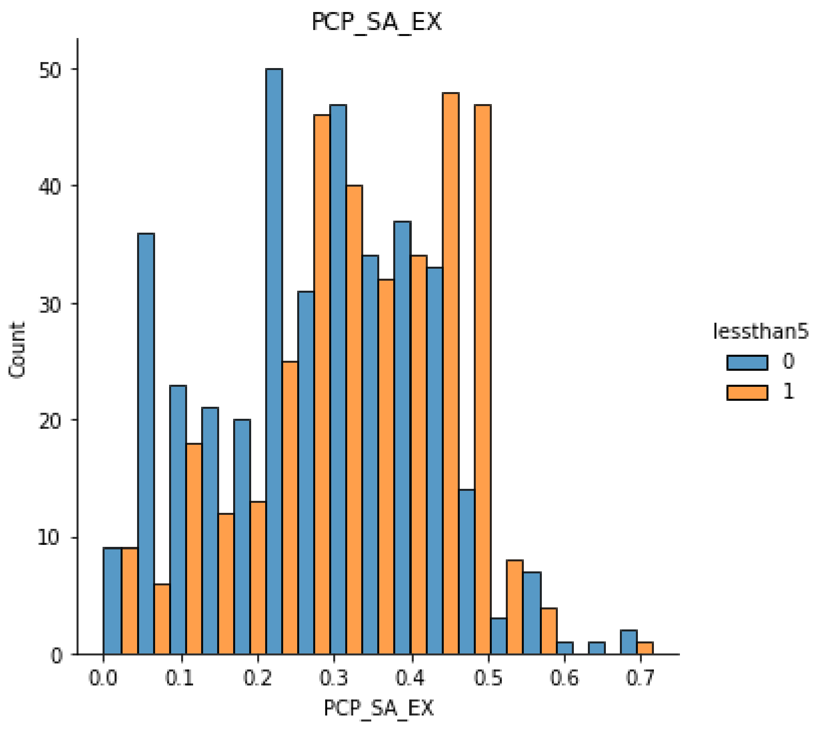

RRI_W
Count
RRI_W
lessthan5
0
1

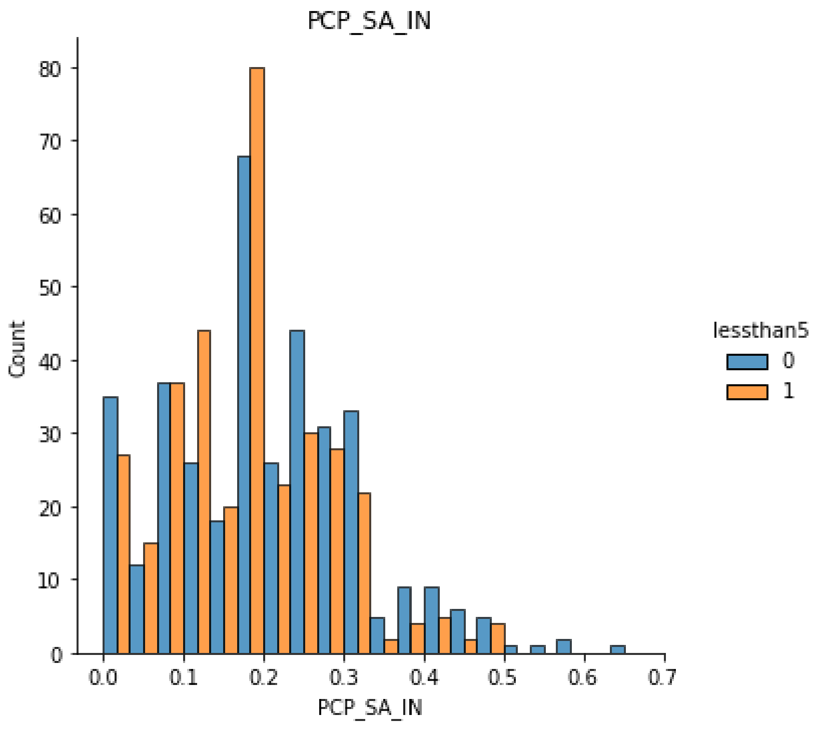

RRI_Y
Count
RRI_Y
lessthan5
0
1
0
50
100
150
200
0.0
0.5
1.0
1.5
2.0


PCP_SS_HE
Count
PCP_SS_HE
lessthan5
0
1
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0


PCP_SS_ST
Count
PCP_SS_ST
lessthan5
0
1
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50
60

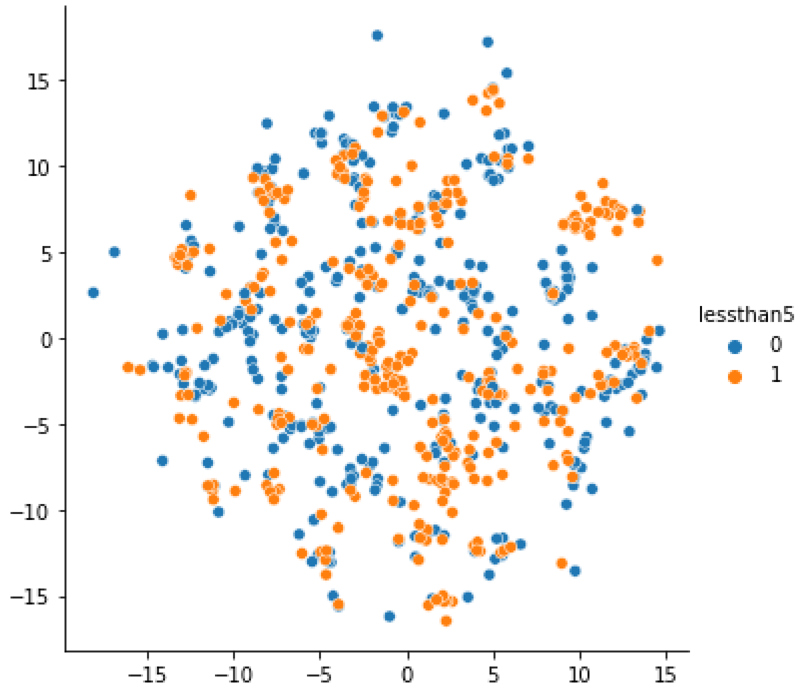

PCP_SA_BU
Count
PCP_SA_BU
lessthan5
0
1
0.2
0.4
0.6
0.8
0
10
20
30
40
50
60





```
In [ ]: from sklearn.preprocessing import StandardScaler
```

In [40]:
```
tsne = TSNE(perplexity=5, n_iter=250)
X = tsne.fit_transform(physicochem_prop)
sns.relplot(x=X[:,0], y=X[:,1], hue=df['lessthan5'])
plt.show()
```



In [ ]:
```

```