Run each cell one after the other.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from matplotlib import *
import sys
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from scipy.stats import bernoulli
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import mutual_info_classif
import pandas as pd
from sklearn.feature_selection import chi2
import numpy as np
from sklearn.ensemble import ExtraTreesClassifier
```

```
dataframe_renewal=pd.read_csv("lease_renewal.csv")
dataframe_renewal.head()
```

```
  lease_id  no_rent_change  rent_change_10  rent_change_20 
lease_length_2  \
0  HPA0001               0               0               0
0
1  HPA0002               0               0               0
0
2  HPA0003               0               0               0
0
3  HPA0004               0               0               0
0
4  HPA0005               0               0               0
0

   lease_length_3  lease_length_1  age_range_under_24  age_range_24_29 
\
0               0               0                   0                0

1               0               0                   0                0

2               0               0                   0                0

3               0               0                   0                0

4               0               0                   0                0


   age_range_30_39  age_range_40_49  age_range_50_59  age_range_60  \
0                0                0                0             0
1                0                0                0             0
2                0                0                0             0
3                0                0                0             0
4                0                0                0             0

   NoFinesViolations  PositiveSurvey  LatePayments  HOA_mandatory
Renewed
0                  0               1             1              0
1
1                  0               0             0              0
0
2                  0               1             0              0
0
3                  0               1             1              0
0
4                  0               0             0              0
0
```

```
dataframe_renewal.describe()
```

```
       no_rent_change  rent_change_10  rent_change_20  lease_length_2
\
count    79850.000000    79850.000000    79850.000000    79850.000000

mean         0.221428        0.023532        0.581866        0.245172

std          0.415210        0.151586        0.493256        0.430192

min          0.000000        0.000000        0.000000        0.000000

25%          0.000000        0.000000        0.000000        0.000000

50%          0.000000        0.000000        1.000000        0.000000

75%          0.000000        0.000000        1.000000        0.000000

max          1.000000        1.000000        1.000000        1.000000


       lease_length_3  lease_length_1  age_range_under_24
age_range_24_29  \
count    79850.000000    79850.000000        79850.000000
79850.000000
mean         0.057495        0.524859            0.038309
0.091947
std          0.232788        0.499385            0.191943
0.288953
min          0.000000        0.000000            0.000000
0.000000
```

```
25%          0.000000        0.000000            0.000000
0.000000
50%          0.000000        1.000000            0.000000
0.000000
75%          0.000000        1.000000            0.000000
0.000000
max          1.000000        1.000000            1.000000
1.000000

       age_range_30_39  age_range_40_49  age_range_50_59  age_range_60
\
count     79850.000000     79850.000000     79850.000000  79850.000000

mean          0.149192         0.108604         0.059136      0.020977

std           0.356280         0.311143         0.235880      0.143308

min           0.000000         0.000000         0.000000      0.000000

25%           0.000000         0.000000         0.000000      0.000000

50%           0.000000         0.000000         0.000000      0.000000

75%           0.000000         0.000000         0.000000      0.000000

max           1.000000         1.000000         1.000000      1.000000


       NoFinesViolations  PositiveSurvey  LatePayments  HOA_mandatory
\
count       79850.000000    79850.000000  79850.000000   79850.000000

mean            0.139249        0.269142      0.566399       0.164133

std             0.346208        0.443517      0.495575       0.370398

min             0.000000        0.000000      0.000000       0.000000

25%             0.000000        0.000000      0.000000       0.000000

50%             0.000000        0.000000      1.000000       0.000000

75%             0.000000        1.000000      1.000000       0.000000

max             1.000000        1.000000      1.000000       1.000000


            Renewed
```

```
count  79850.000000
mean       0.195892
std        0.396888
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        1.000000
```

```
dataframe_renewal.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 79850 entries, 0 to 79849
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   lease_id            79850 non-null  object
 1   no_rent_change      79850 non-null  int64 
 2   rent_change_10      79850 non-null  int64 
 3   rent_change_20      79850 non-null  int64 
 4   lease_length_2      79850 non-null  int64 
 5   lease_length_3      79850 non-null  int64 
 6   lease_length_1      79850 non-null  int64 
 7   age_range_under_24  79850 non-null  int64 
 8   age_range_24_29     79850 non-null  int64 
 9   age_range_30_39     79850 non-null  int64 
 10  age_range_40_49     79850 non-null  int64 
 11  age_range_50_59     79850 non-null  int64 
 12  age_range_60        79850 non-null  int64 
 13  NoFinesViolations   79850 non-null  int64 
 14  PositiveSurvey      79850 non-null  int64 
 15  LatePayments        79850 non-null  int64 
 16  HOA_mandatory       79850 non-null  int64 
 17  Renewed             79850 non-null  int64 
dtypes: int64(17), object(1)
memory usage: 11.0+ MB
```

```
import pandas_profiling
```

```
dataframe_renewal.profile_report()
```

{"version_major":2,"version_minor":0,"model_id":"8346ea3721d04578b6dc45e92388cc7d"}

{"version_major":2,"version_minor":0,"model_id":"65df5a7935c64a5a9b20957d84c5bdf5"}

{"version_major":2,"version_minor":0,"model_id":"e58cbcf8c691458097a294976b4b213f"}

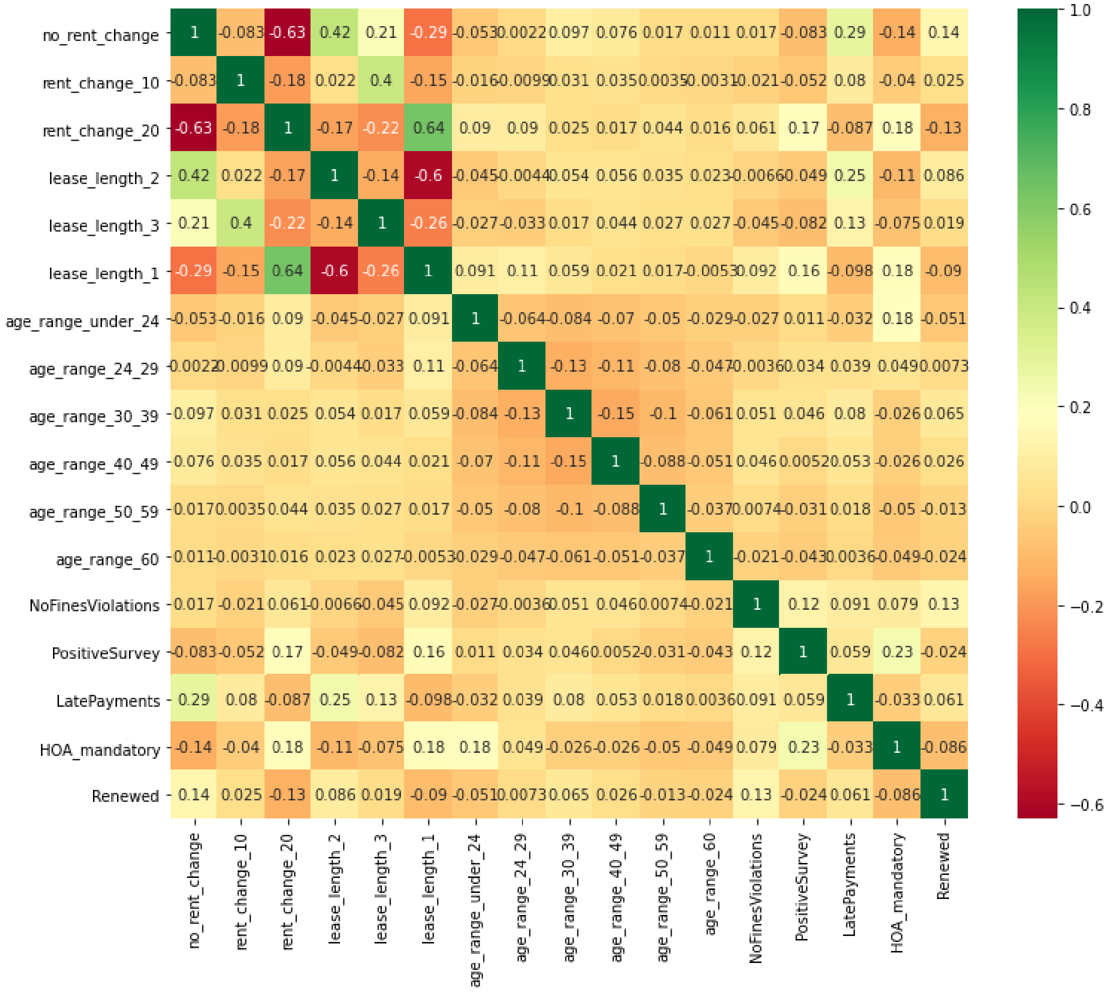
```
<IPython.core.display.HTML object>
```

```
dataframe_renewal.isnull().sum()
```

```
lease_id               0
no_rent_change         0
rent_change_10         0
rent_change_20         0
lease_length_2         0
lease_length_3         0
lease_length_1         0
age_range_under_24     0
age_range_24_29        0
age_range_30_39        0
age_range_40_49        0
age_range_50_59        0
age_range_60           0
NoFinesViolations      0
PositiveSurvey         0
LatePayments           0
HOA_mandatory          0
Renewed                0
dtype: int64
```

HeatMap visualization

```
plt.figure(figsize=(12,10))  # on this line I just set the size of
figure to 12 by 10.
p=sns.heatmap(dataframe_renewal.corr(), annot=True,cmap ='RdYlGn')  #
seaborn has very simple solution for heatmap
```


no_rent_change 1 -0.083 -0.63 0.42 0.21 -0.29 -0.053 0.0022 0.097 0.076 0.017 0.011 0.017 -0.083 0.29 -0.14 0.14
rent_change_10 -0.083 1 -0.18 0.022 0.4 -0.15 -0.016 -0.0099 0.031 0.035 0.0035 -0.0031 -0.021 -0.052 0.08 -0.04 0.025
rent_change_20 -0.63 -0.18 1 -0.17 -0.22 0.64 0.09 0.09 0.025 0.017 0.044 0.016 0.061 0.17 -0.087 0.18 -0.13
lease_length_2 0.42 0.022 -0.17 1 -0.14 -0.6 -0.045 -0.0044 0.054 0.056 0.035 0.023 -0.0066 -0.049 0.25 -0.11 0.086
lease_length_3 0.21 0.4 -0.22 -0.14 1 -0.26 -0.027 -0.033 0.017 0.044 0.027 0.027 -0.045 -0.082 0.13 -0.075 0.019
lease_length_1 -0.29 -0.15 0.64 -0.6 -0.26 1 0.091 0.11 0.059 0.021 0.017 -0.0053 0.092 0.16 -0.098 0.18 -0.09
age_range_under_24 -0.053 -0.016 0.09 -0.045 -0.027 0.091 1 -0.064 -0.084 -0.07 -0.05 -0.029 -0.027 0.011 -0.032 0.18 -0.051
age_range_24_29 0.0022 -0.0099 0.09 -0.0044 -0.033 0.11 -0.064 1 -0.13 -0.11 -0.08 -0.047 -0.0036 0.034 0.039 0.049 0.0073
age_range_30_39 0.097 0.031 0.025 0.054 0.017 0.059 -0.084 -0.13 1 -0.15 -0.1 -0.061 0.051 0.046 0.08 -0.026 0.065
age_range_40_49 0.076 0.035 0.017 0.056 0.044 0.021 -0.07 -0.11 -0.15 1 -0.088 -0.051 0.046 0.0052 0.053 -0.026 0.026
age_range_50_59 0.017 0.0035 0.044 0.035 0.027 0.017 -0.05 -0.08 -0.1 -0.088 1 -0.037 0.0074 -0.031 0.018 -0.05 -0.013
age_range_60 0.011 -0.0031 0.016 0.023 0.027 -0.0053 -0.029 -0.047 -0.061 -0.051 -0.037 1 -0.021 -0.043 0.0036 -0.049 -0.024
NoFinesViolations 0.017 -0.021 0.061 -0.0066 -0.045 0.092 -0.027 -0.0036 0.051 0.046 0.0074 -0.021 1 0.12 0.091 0.079 0.13
PositiveSurvey -0.083 -0.052 0.17 -0.049 -0.082 0.16 0.011 0.034 0.046 0.0052 -0.031 -0.043 0.12 1 0.059 0.23 -0.024
LatePayments 0.29 0.08 -0.087 0.25 0.13 -0.098 -0.032 0.039 0.08 0.053 0.018 0.0036 0.091 0.059 1 -0.033 0.061
HOA_mandatory -0.14 -0.04 0.18 -0.11 -0.075 0.18 0.18 0.049 -0.026 -0.026 -0.05 -0.049 0.079 0.23 -0.033 1 -0.086
Renewed 0.14 0.025 -0.13 0.086 0.019 -0.09 -0.051 0.0073 0.065 0.026 -0.013 -0.024 0.13 -0.024 0.061 -0.086 1
no_rent_change
rent_change_10
rent_change_20
lease_length_2
lease_length_3
lease_length_1
age_range_under_24
age_range_24_29
age_range_30_39
age_range_40_49
age_range_50_59
age_range_60
NoFinesViolations
PositiveSurvey
LatePayments
HOA_mandatory
Renewed
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Observations:

1) Those who were on their lease for first term were most likely had an increase of 20 ppercentage of rent

2) Those who were on their lease for second term were most liekly had no increase in their rent.

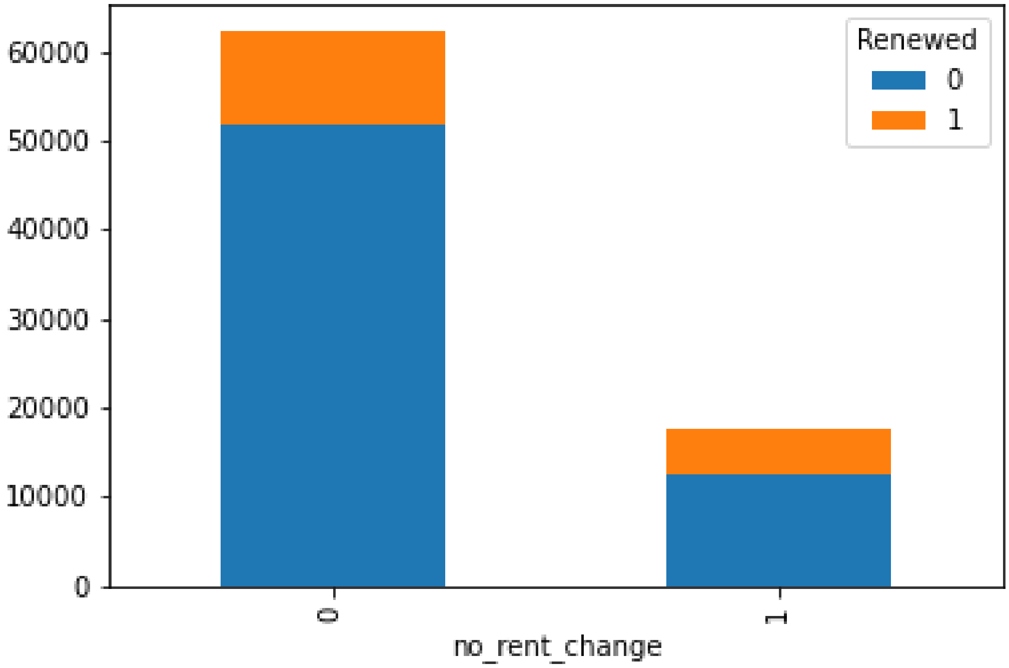
3) Those who were on their lease for third term were most likely had an increase of 10 percenatge of rent.

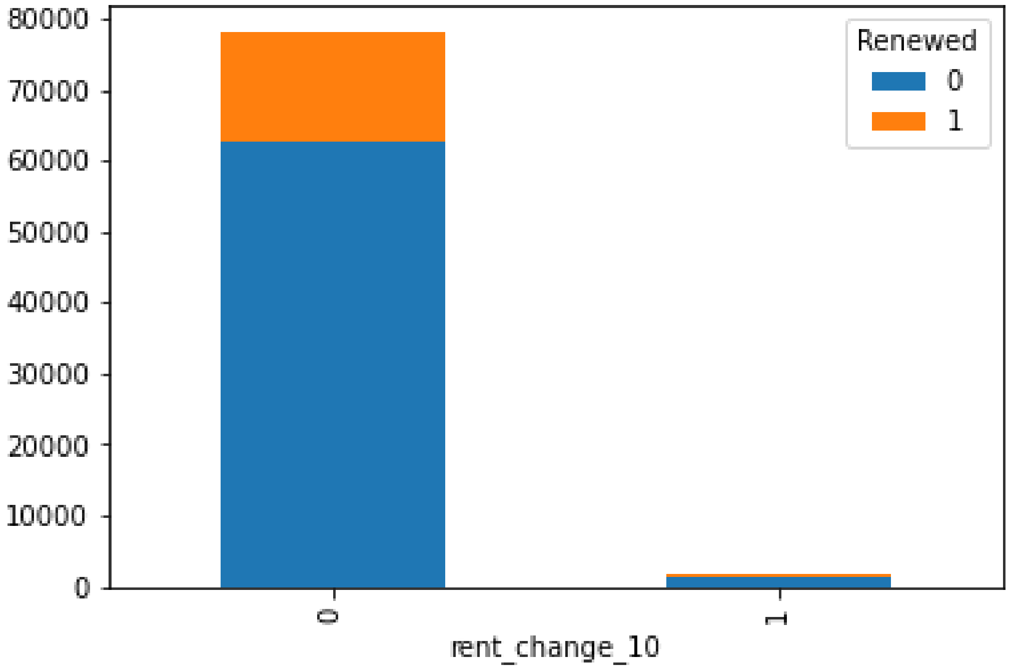
Bar graph to show the count of every value of each independent features to that of the dependent feature.

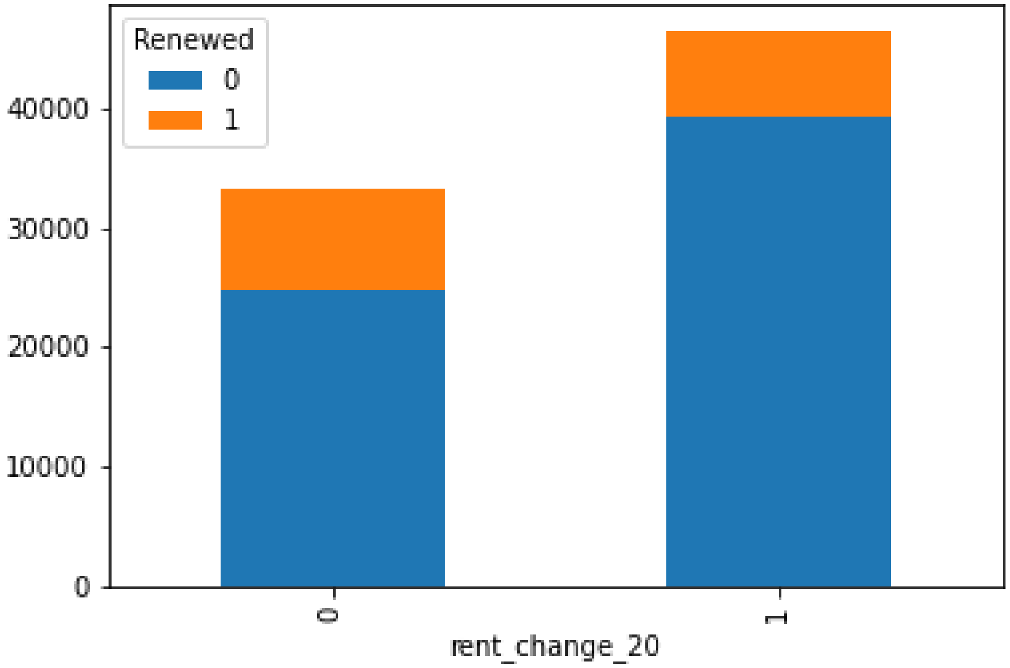
```
for x in range(1,len(dataframe_renewal.columns)-1):

pd.crosstab(dataframe_renewal.iloc[:,x],dataframe_renewal['Renewed']).
plot(kind="bar",stacked=True)
    print(dataframe_renewal.iloc[:,x].value_counts())
```

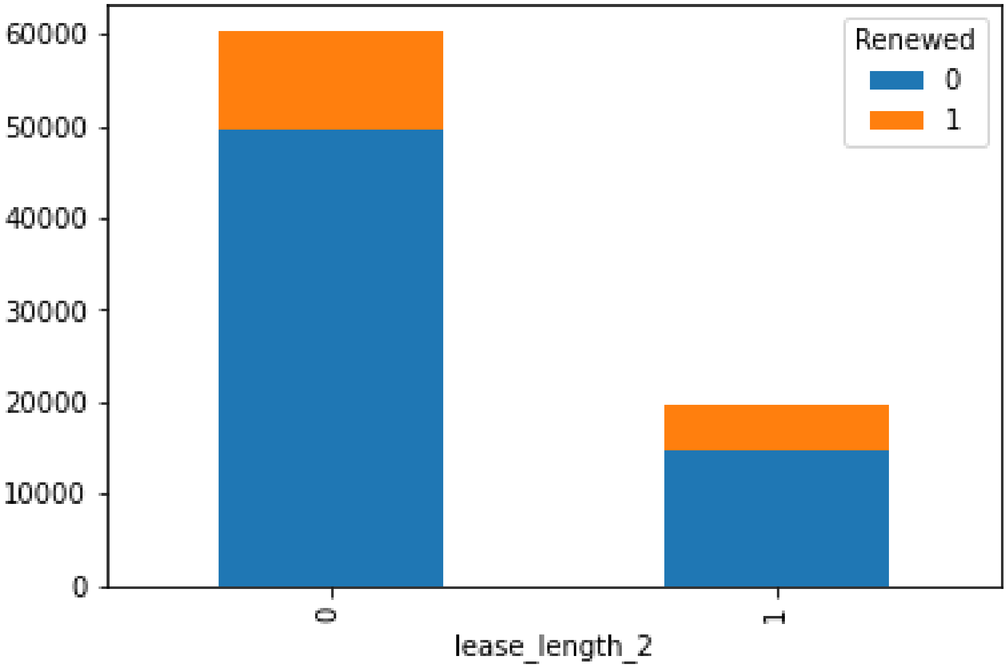
```
0    62169
1    17681
Name: no_rent_change, dtype: int64
0    77971
1     1879
Name: rent_change_10, dtype: int64
1    46462
0    33388
Name: rent_change_20, dtype: int64
0    60273
1    19577
Name: lease_length_2, dtype: int64
0    75259
1     4591
Name: lease_length_3, dtype: int64
1    41910
0    37940
Name: lease_length_1, dtype: int64
0    76791
1     3059
Name: age_range_under_24, dtype: int64
0    72508
1     7342
```

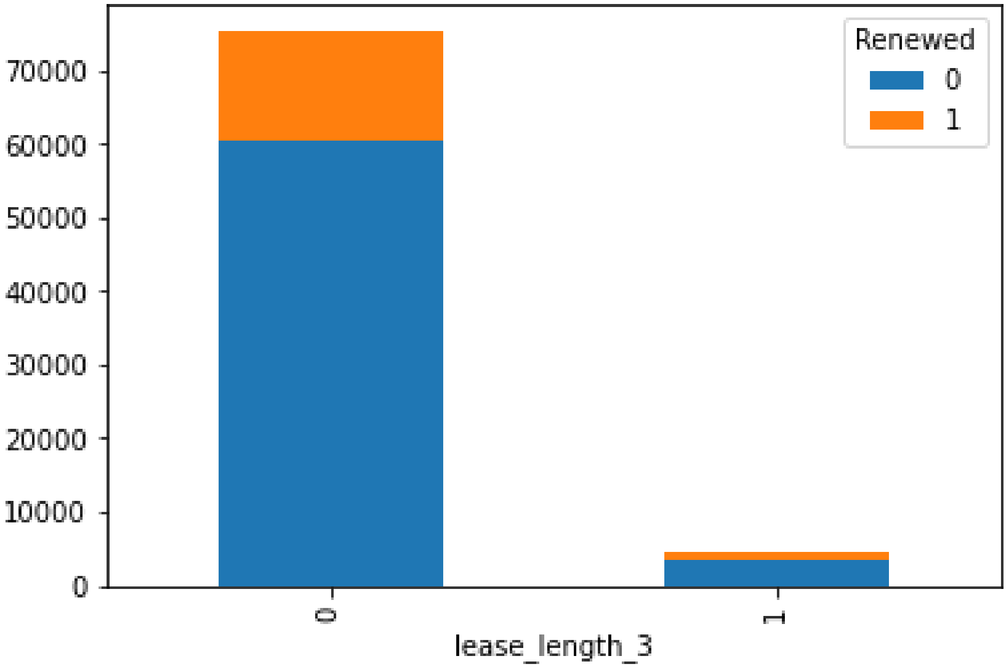
```
Name: age_range_24_29, dtype: int64
0    67937
1    11913
Name: age_range_30_39, dtype: int64
0    71178
1     8672
Name: age_range_40_49, dtype: int64
0    75128
1     4722
Name: age_range_50_59, dtype: int64
0    78175
1     1675
Name: age_range_60, dtype: int64
0    68731
1    11119
Name: NoFinesViolations, dtype: int64
0    58359
1    21491
Name: PositiveSurvey, dtype: int64
1    45227
0    34623
Name: LatePayments, dtype: int64
0    66744
1    13106
Name: HOA_mandatory, dtype: int64
```

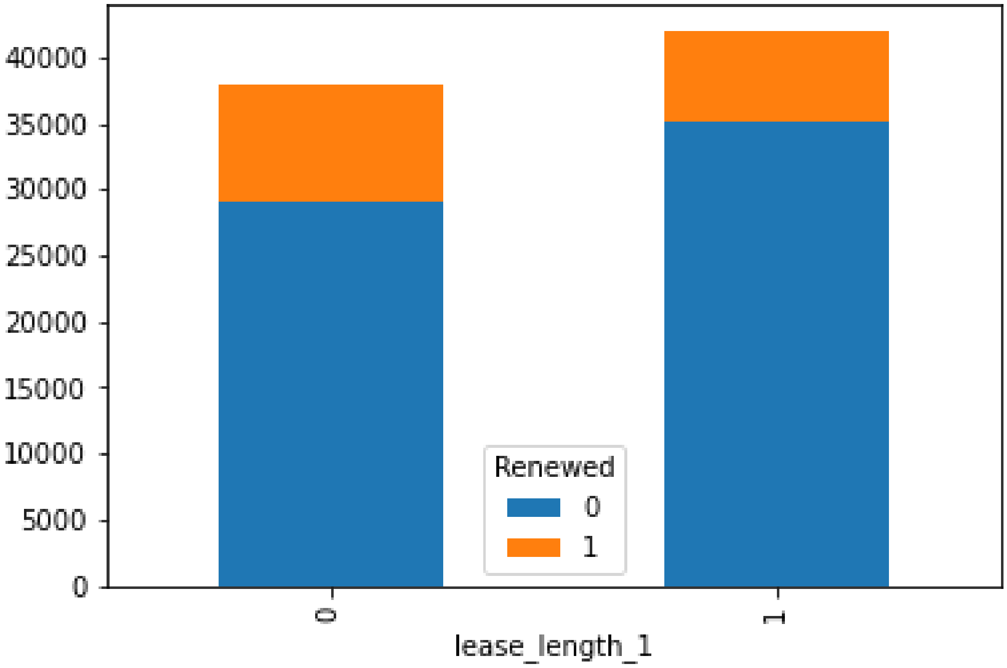


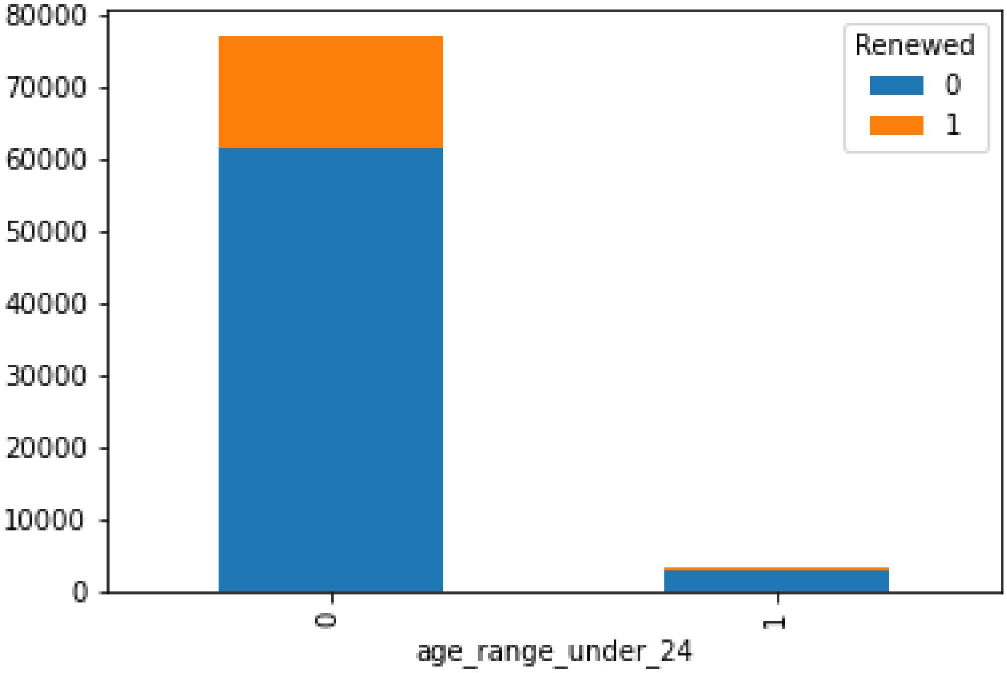

80000
70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
rent_change_10

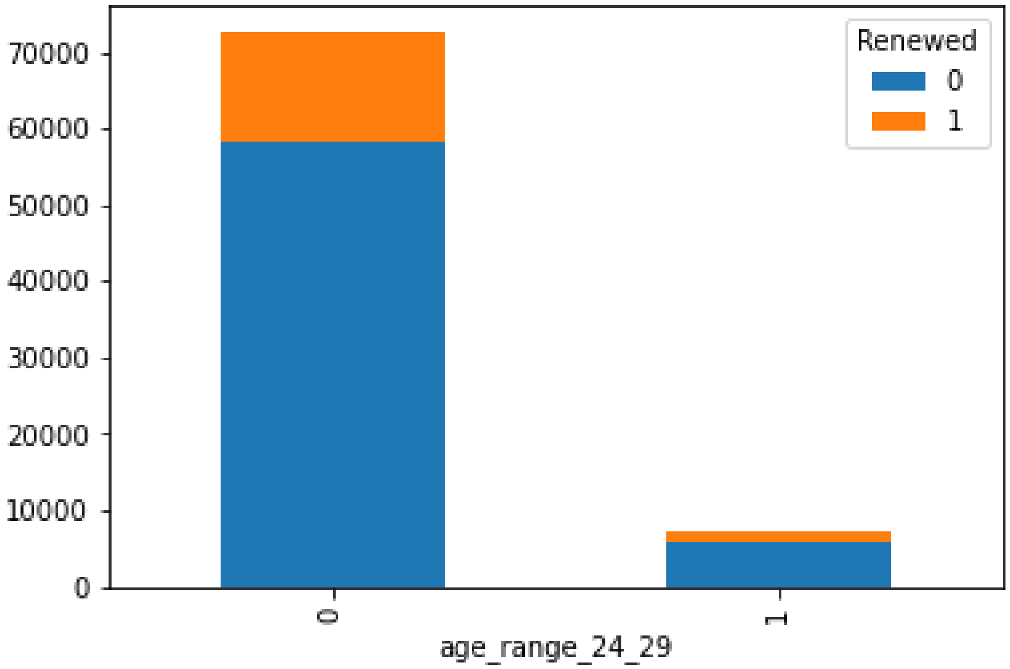

40000
30000
20000
10000
0
Renewed
0
1
0
1
rent_change_20

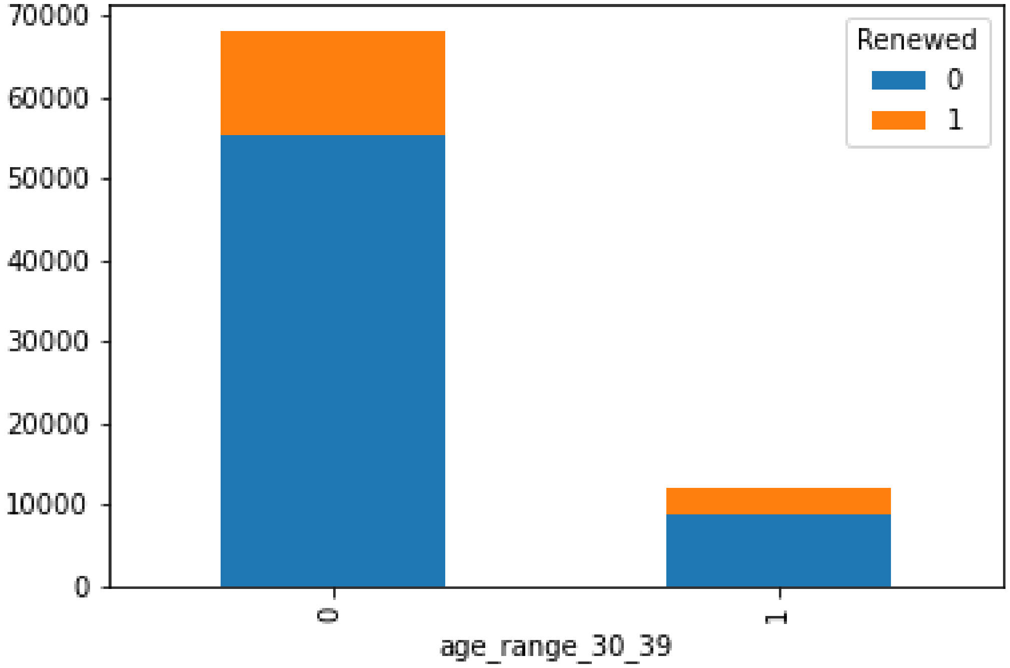

60000
50000
40000
30000
20000
10000
0
0
1
lease_length_2
Renewed
0
1

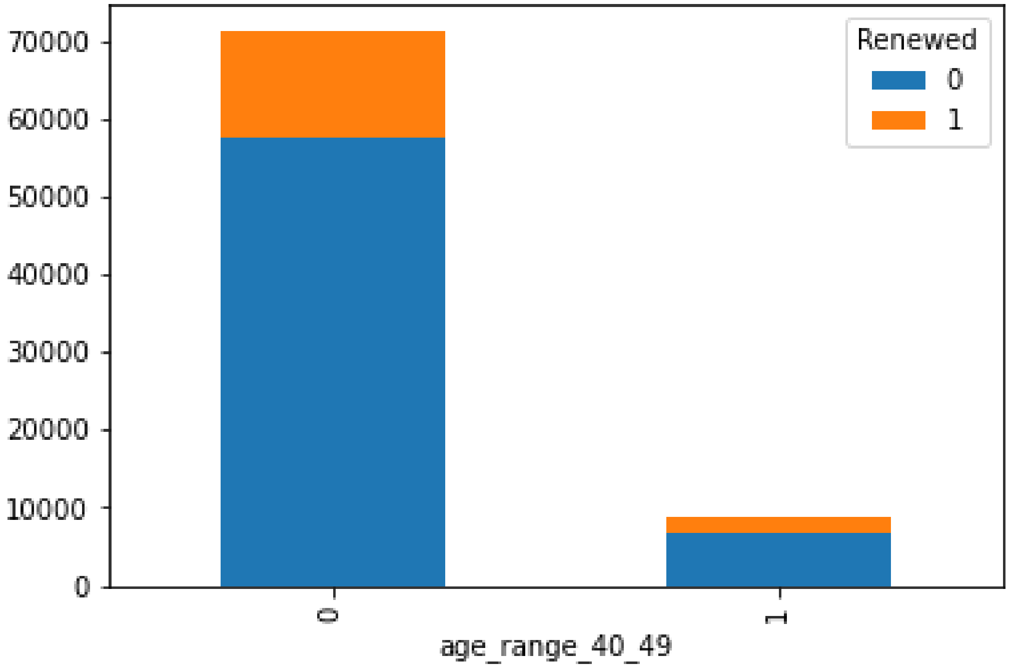

70000
60000
50000
40000
30000
20000
10000
0
0
1
lease_length_3
Renewed
0
1

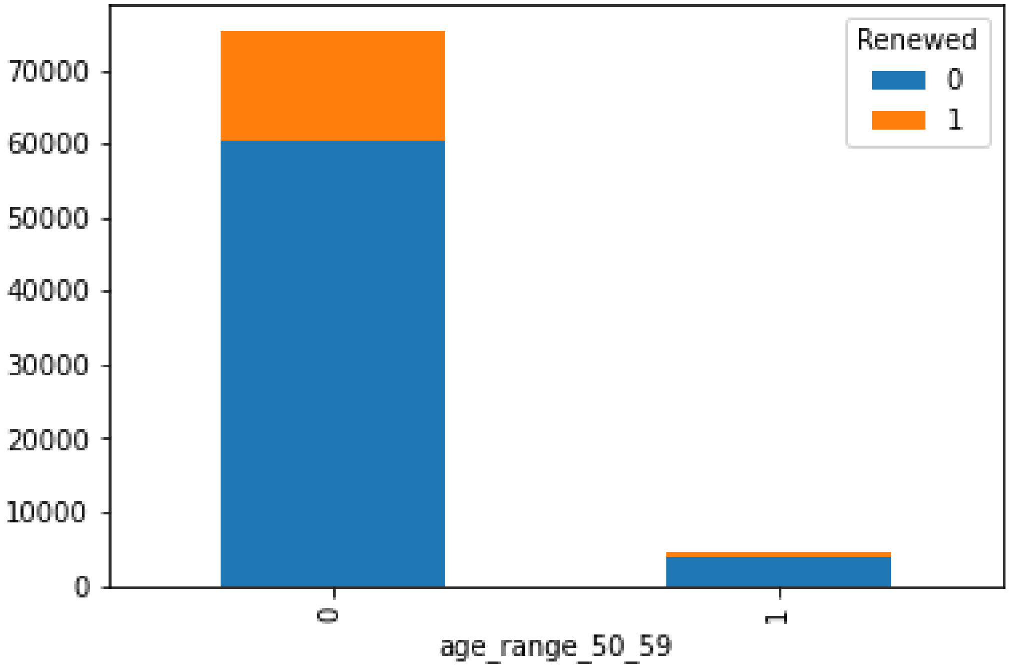

40000
35000
30000
25000
20000
15000
10000
5000
0
Renewed
0
1
0
1
lease_length_1

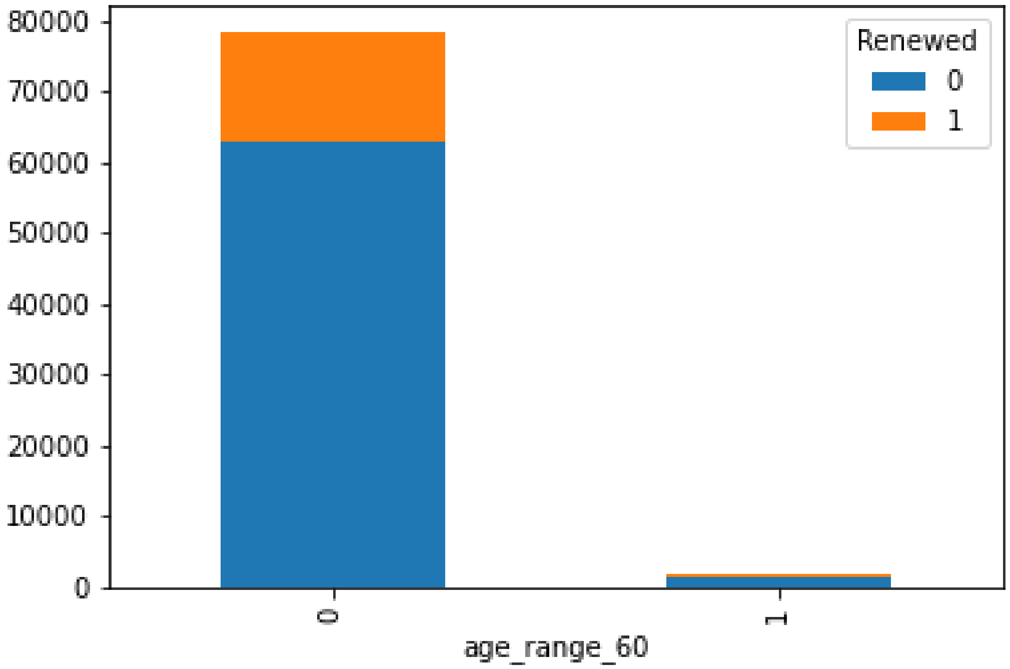

80000
70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
age_range_under_24

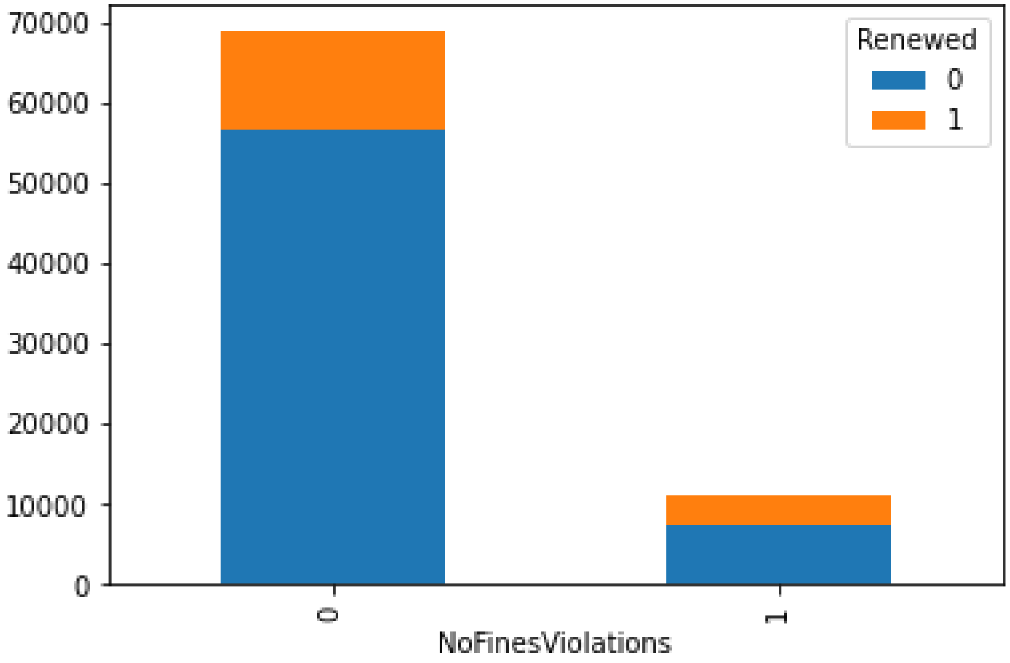

70000
60000
50000
40000
30000
20000
10000
0
0
1
age_range_24_29
Renewed
0
1

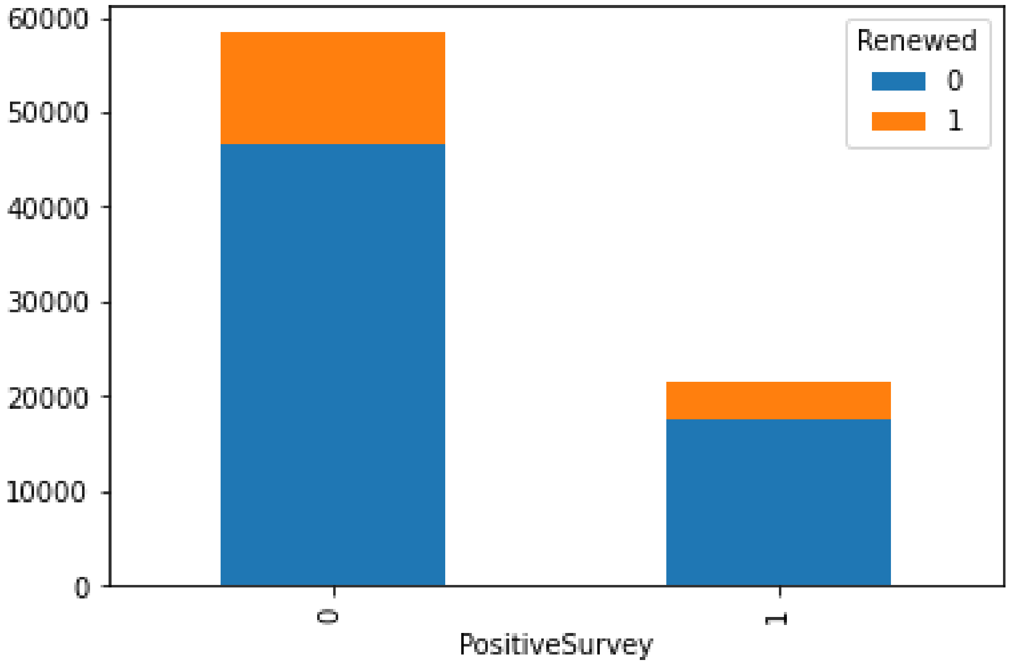

70000
60000
50000
40000
30000
20000
10000
0
0
1
age_range_30_39
Renewed
0
1

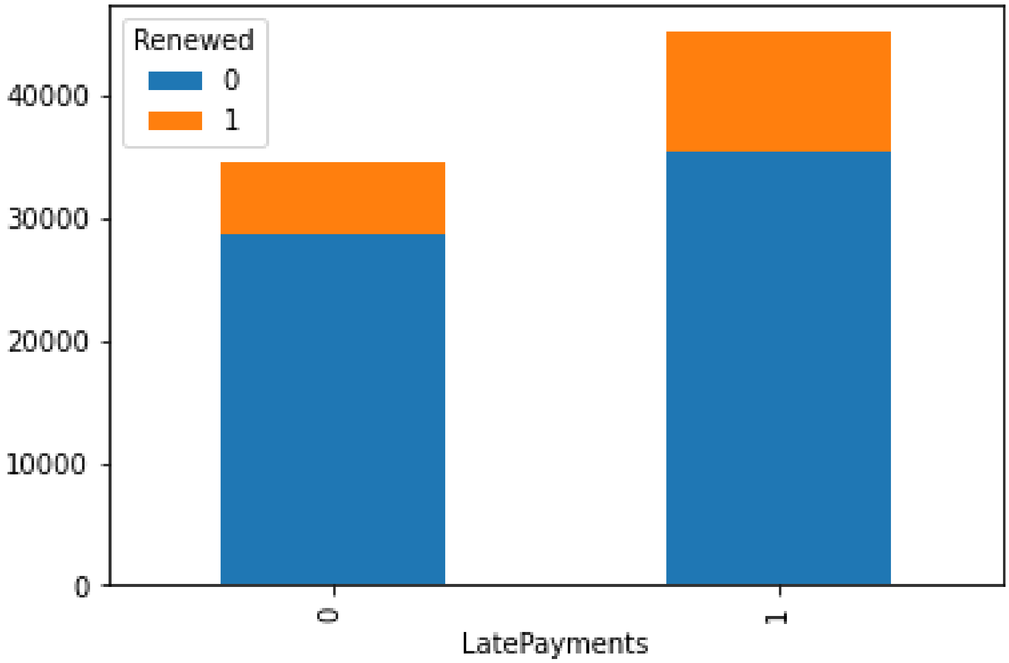

70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
age_range_40_49

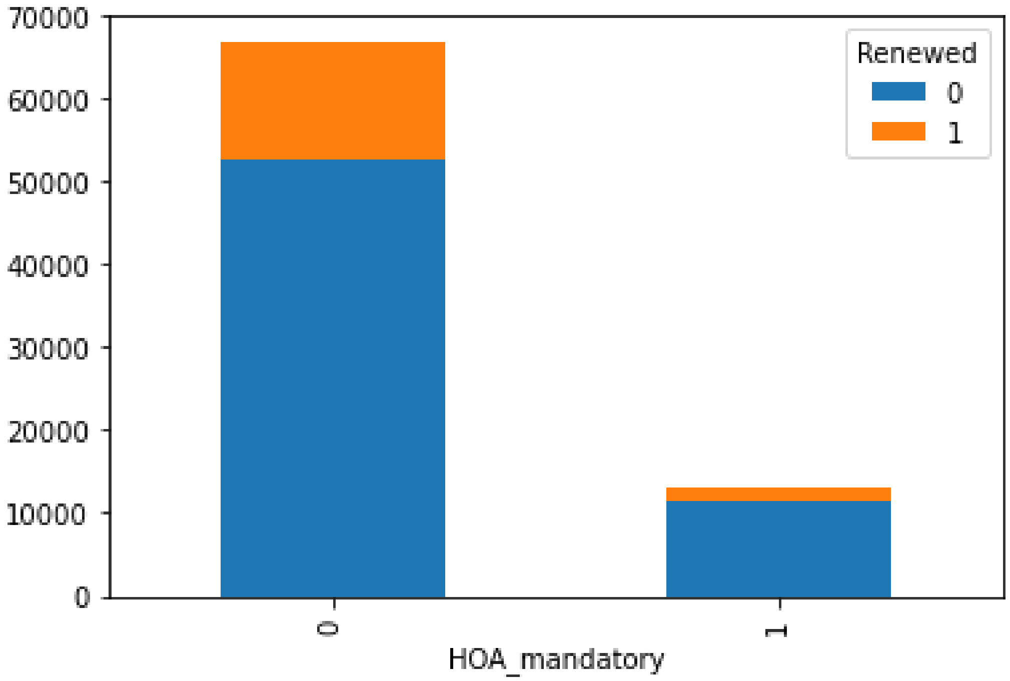

70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
age_range_50_59


80000
70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
age_range_60


70000
60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
NoFinesViolations

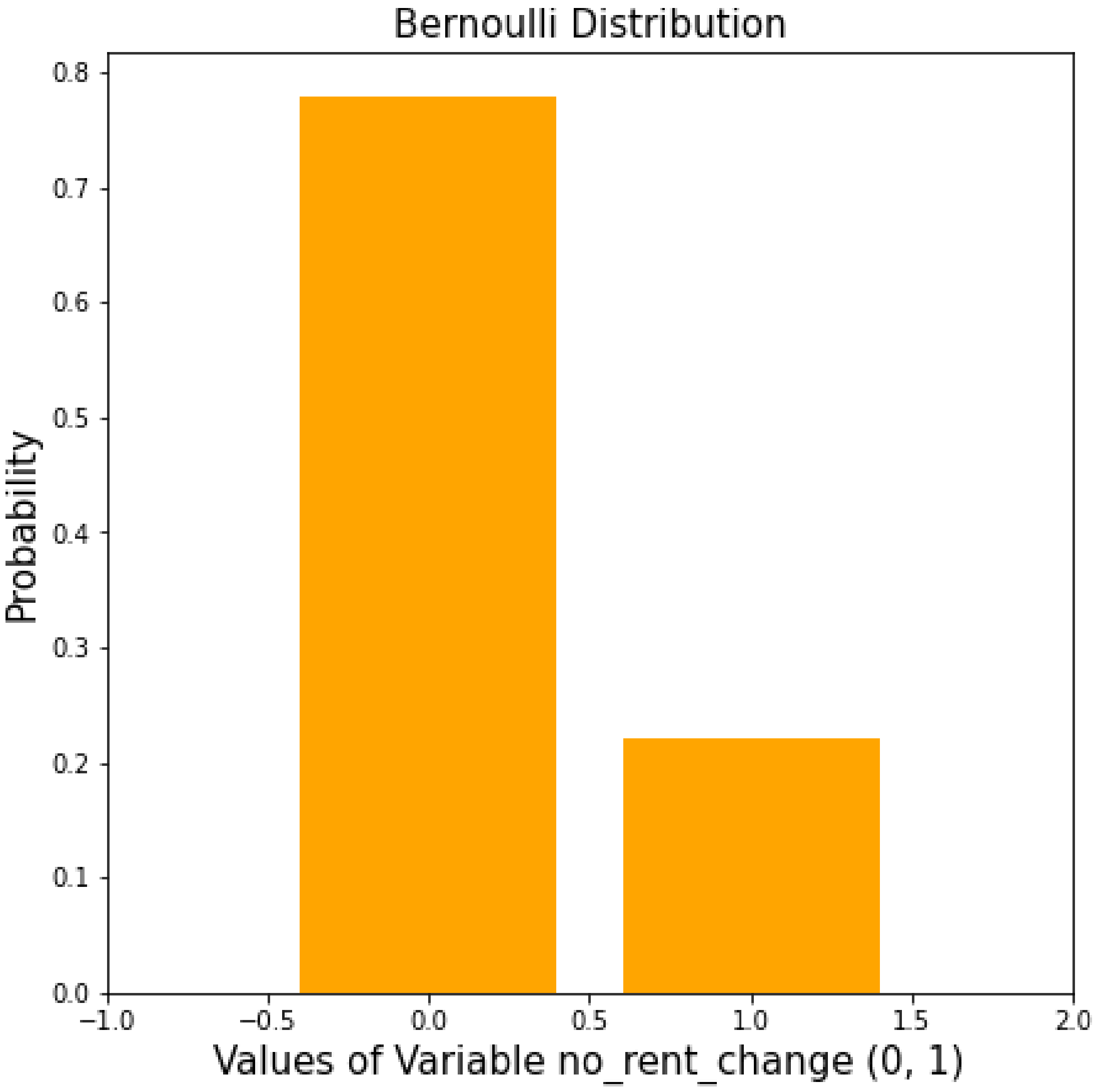

60000
50000
40000
30000
20000
10000
0
Renewed
0
1
0
1
PositiveSurvey

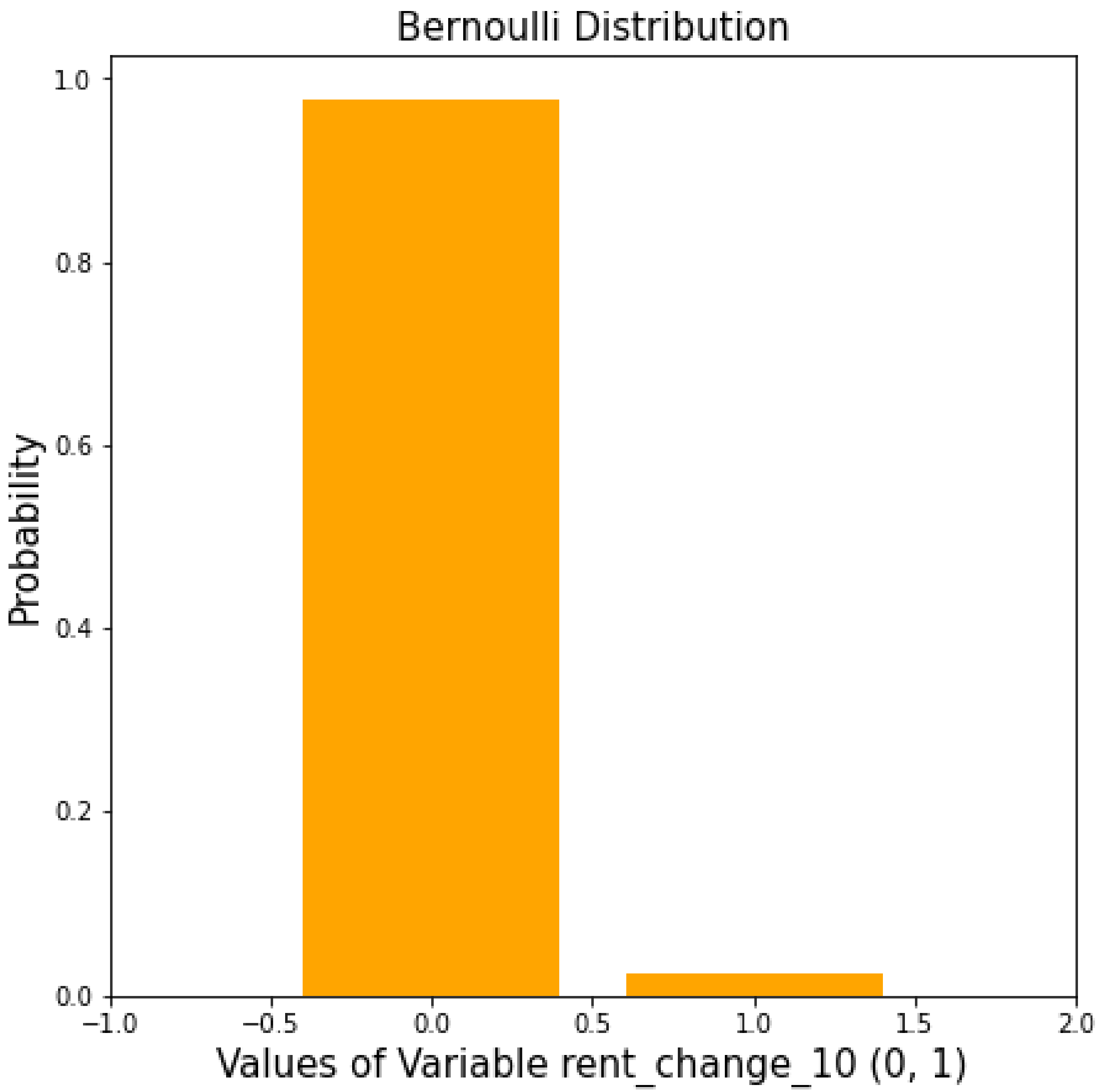

40000
30000
20000
10000
0
Renewed
0
1
0
1
LatePayments

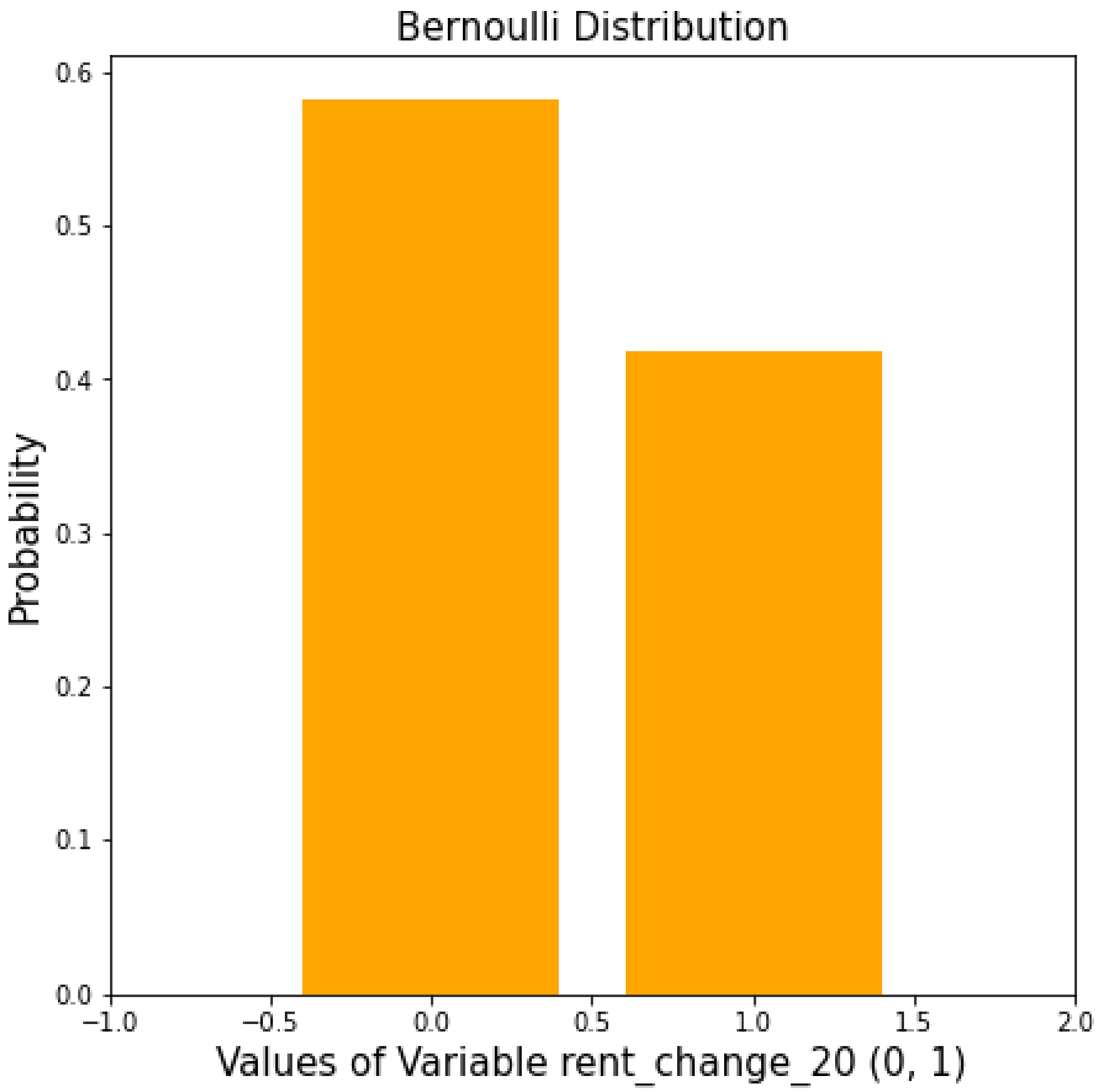


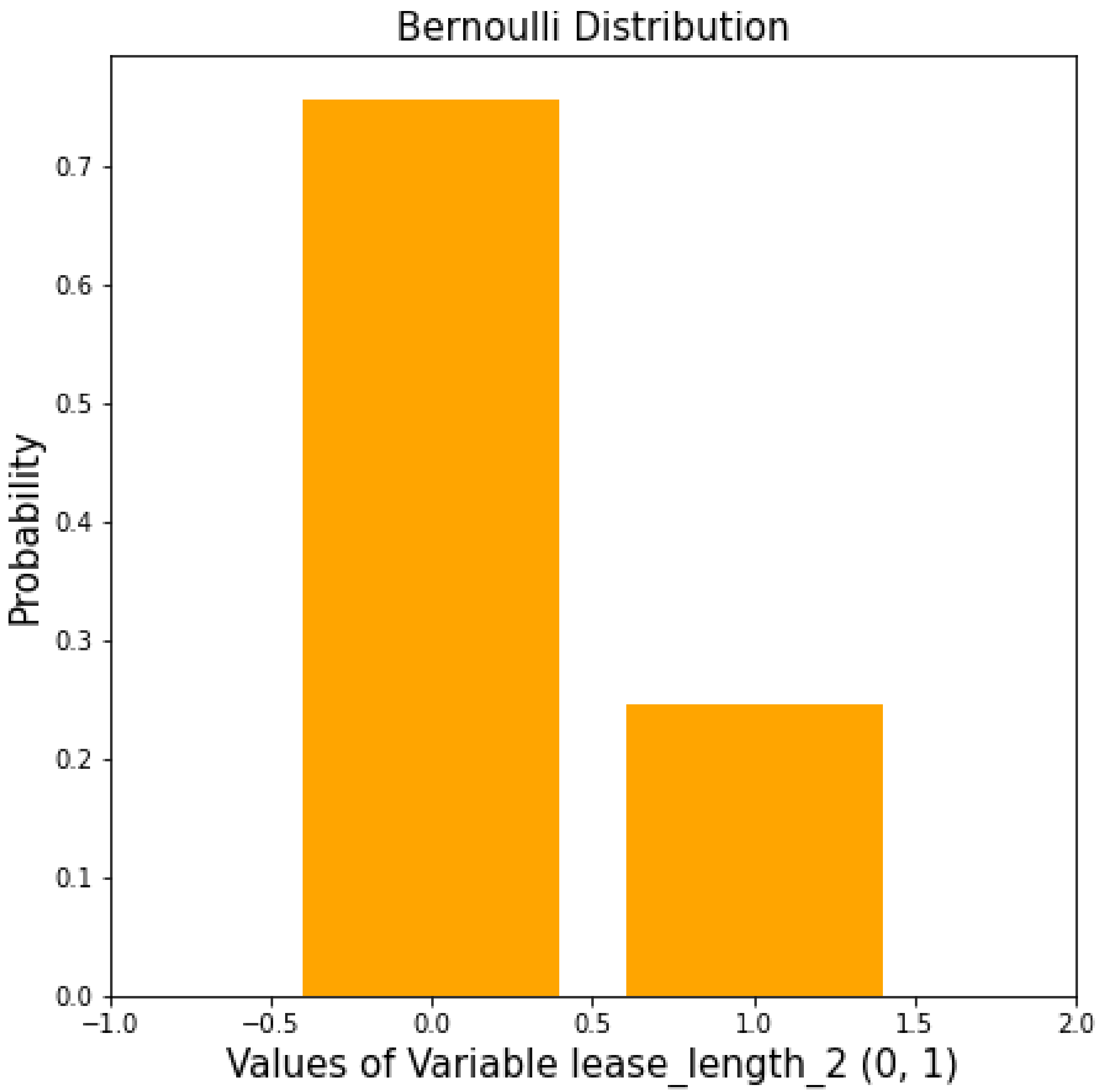
plotting the probability of each features of 0's and 1's.

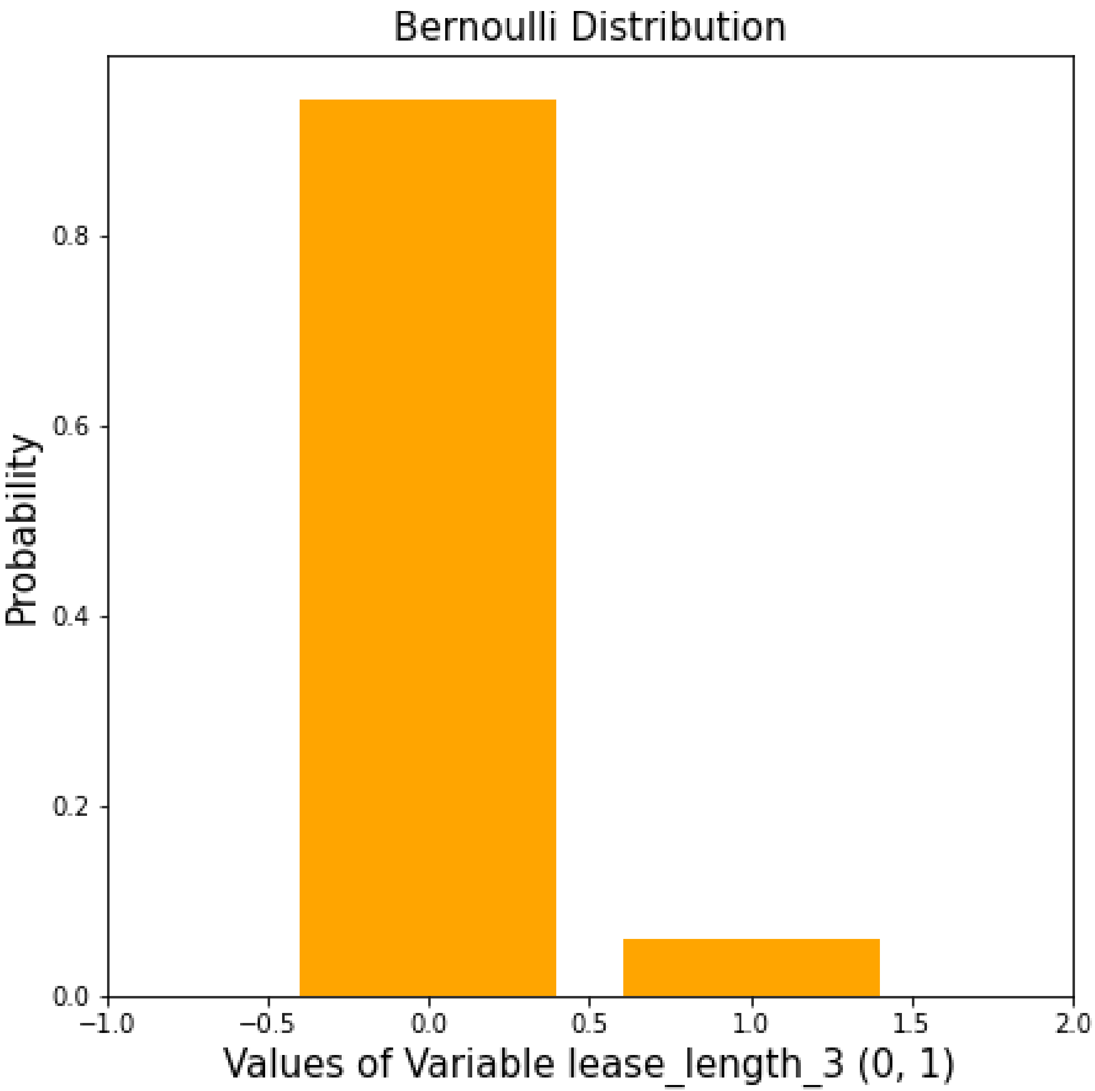
```
for x in dataframe_renewal.columns[1:]:
    val=dataframe_renewal[x].value_counts().to_list()[1]
    bd=bernoulli(val/79850)
    uniqueval= dataframe_renewal[x].unique().tolist()
    plt.figure(figsize=(7,7))
    plt.xlim(-1, 2)
    plt.bar(uniqueval, bd.pmf(uniqueval), color='orange')
    plt.title('Bernoulli Distribution', fontsize='15')
    plt.xlabel('Values of Variable '+x+' (0, 1)', fontsize='15')
    plt.ylabel('Probability', fontsize='15')
    plt.show()
```

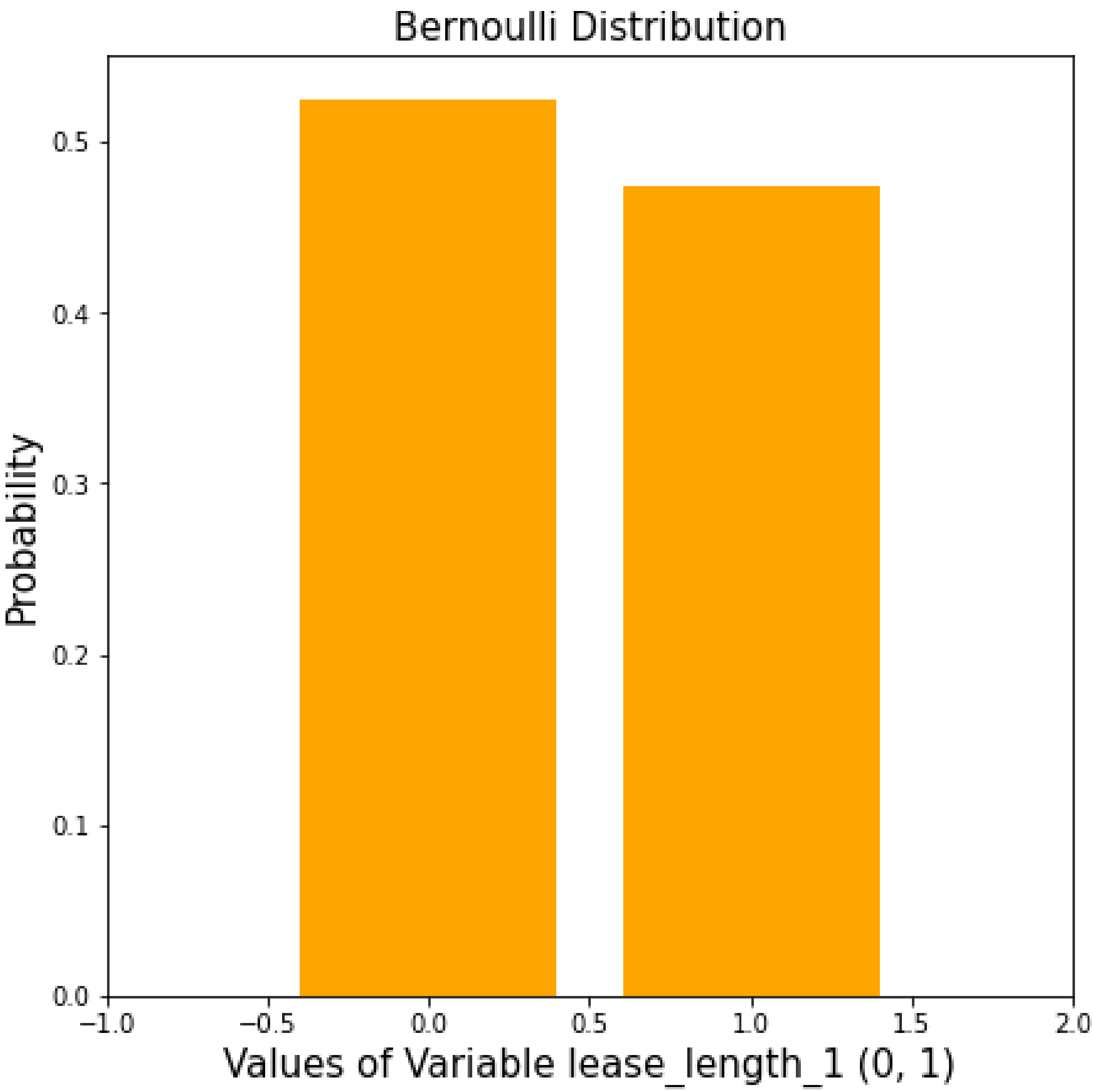

Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable no_rent_change (0, 1)

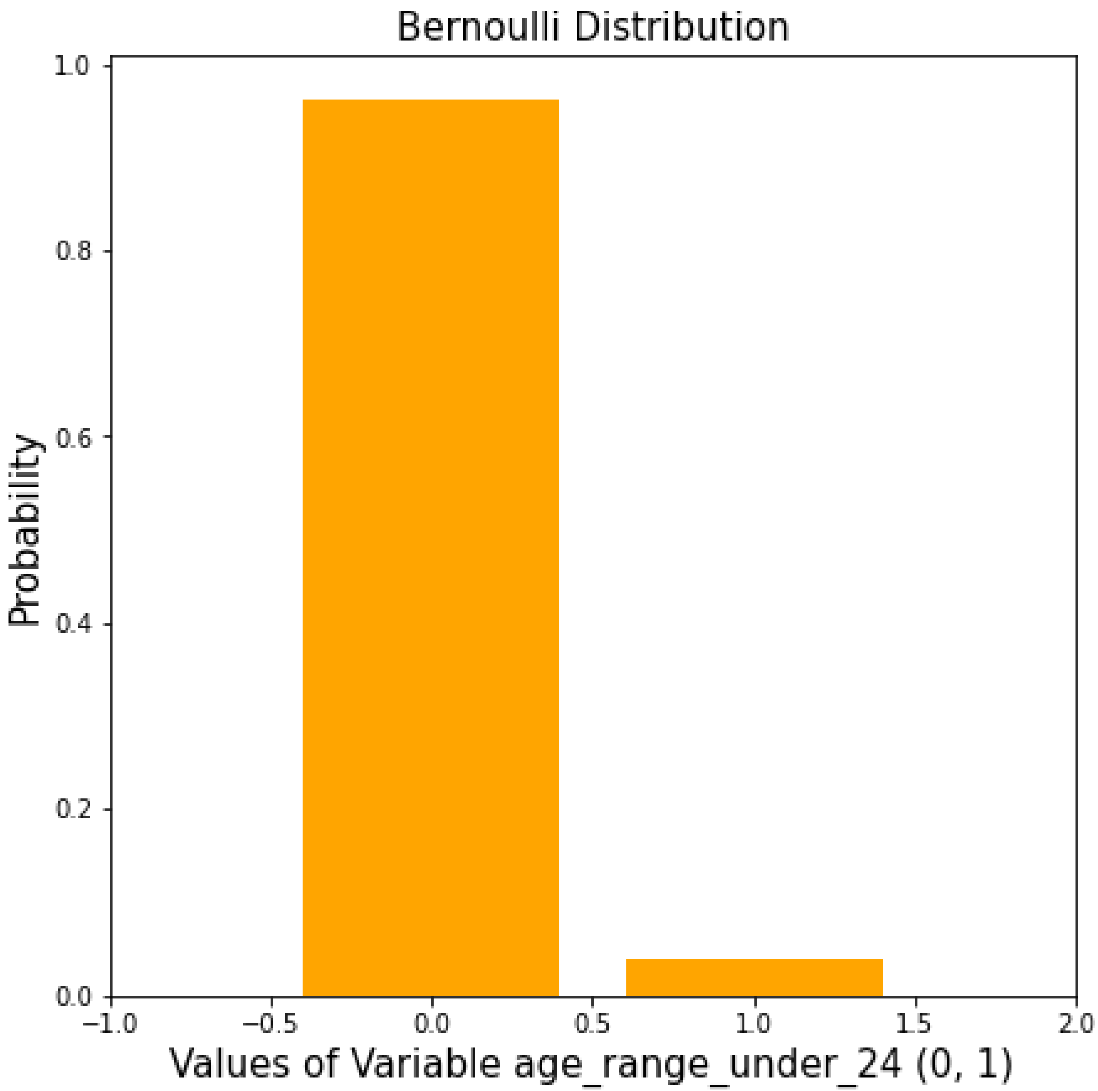

Bernoulli Distribution
Probability
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable rent_change_10 (0, 1)

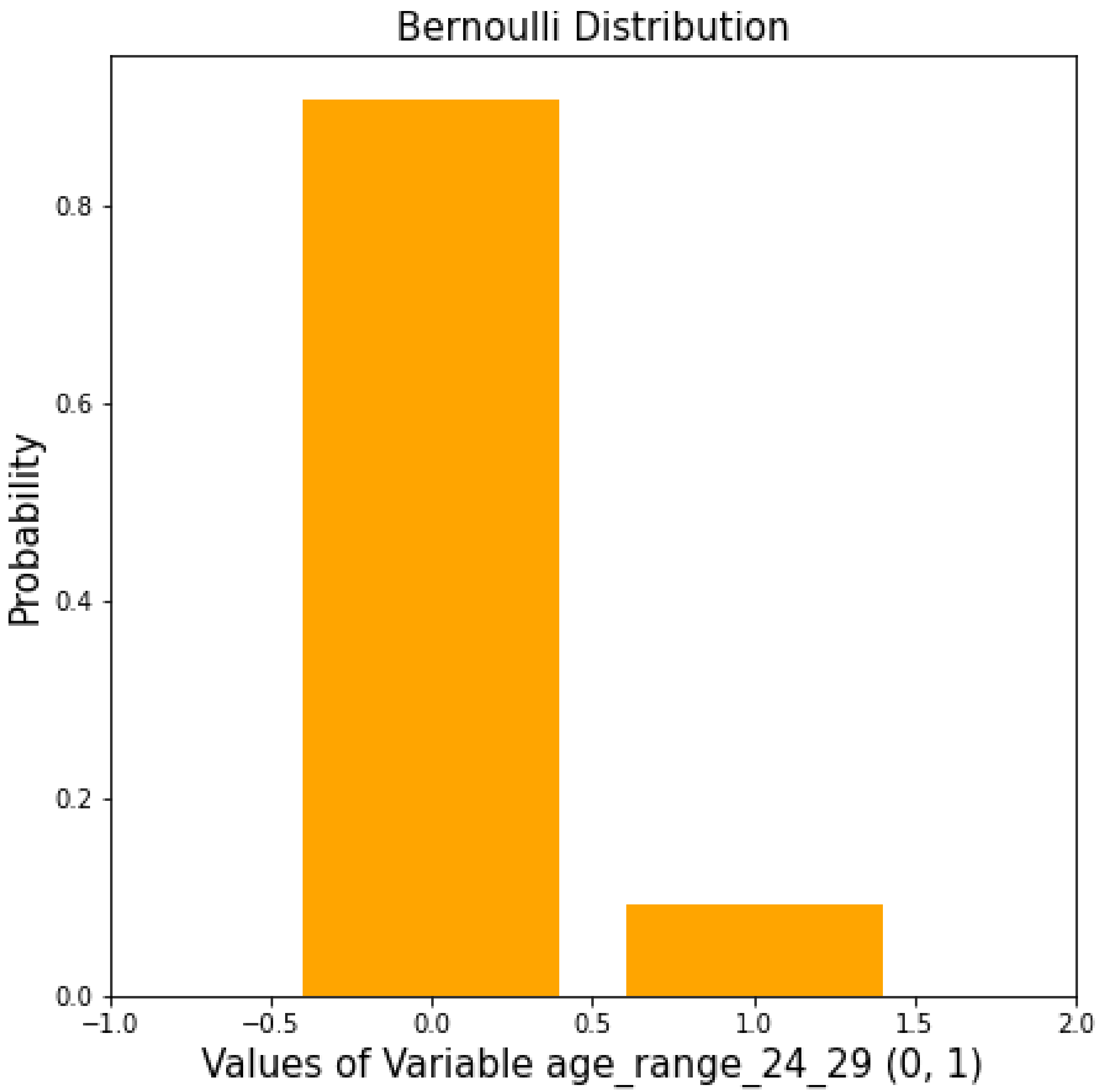

Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable rent_change_20 (0, 1)

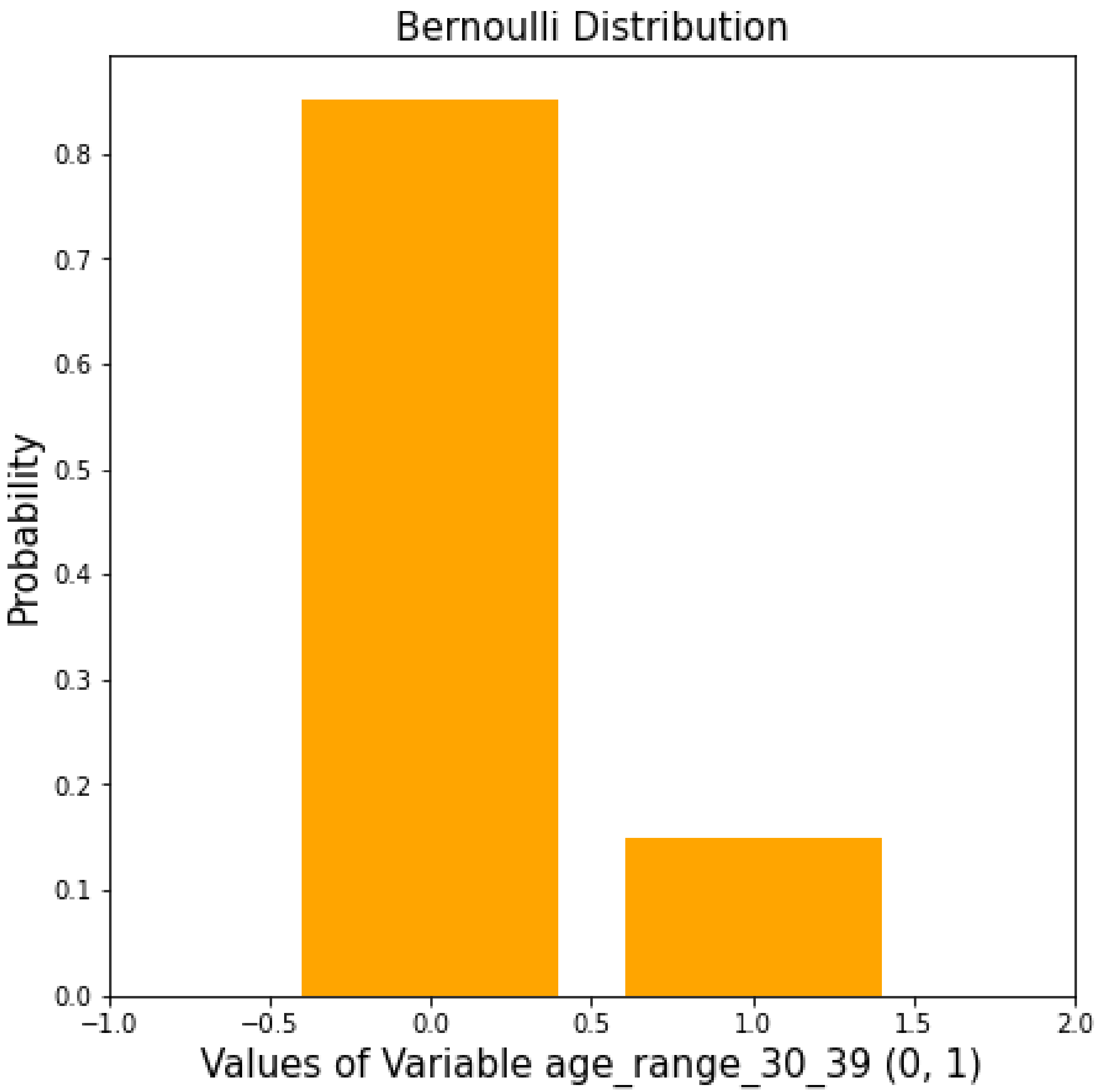

Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable lease_length_2 (0, 1)

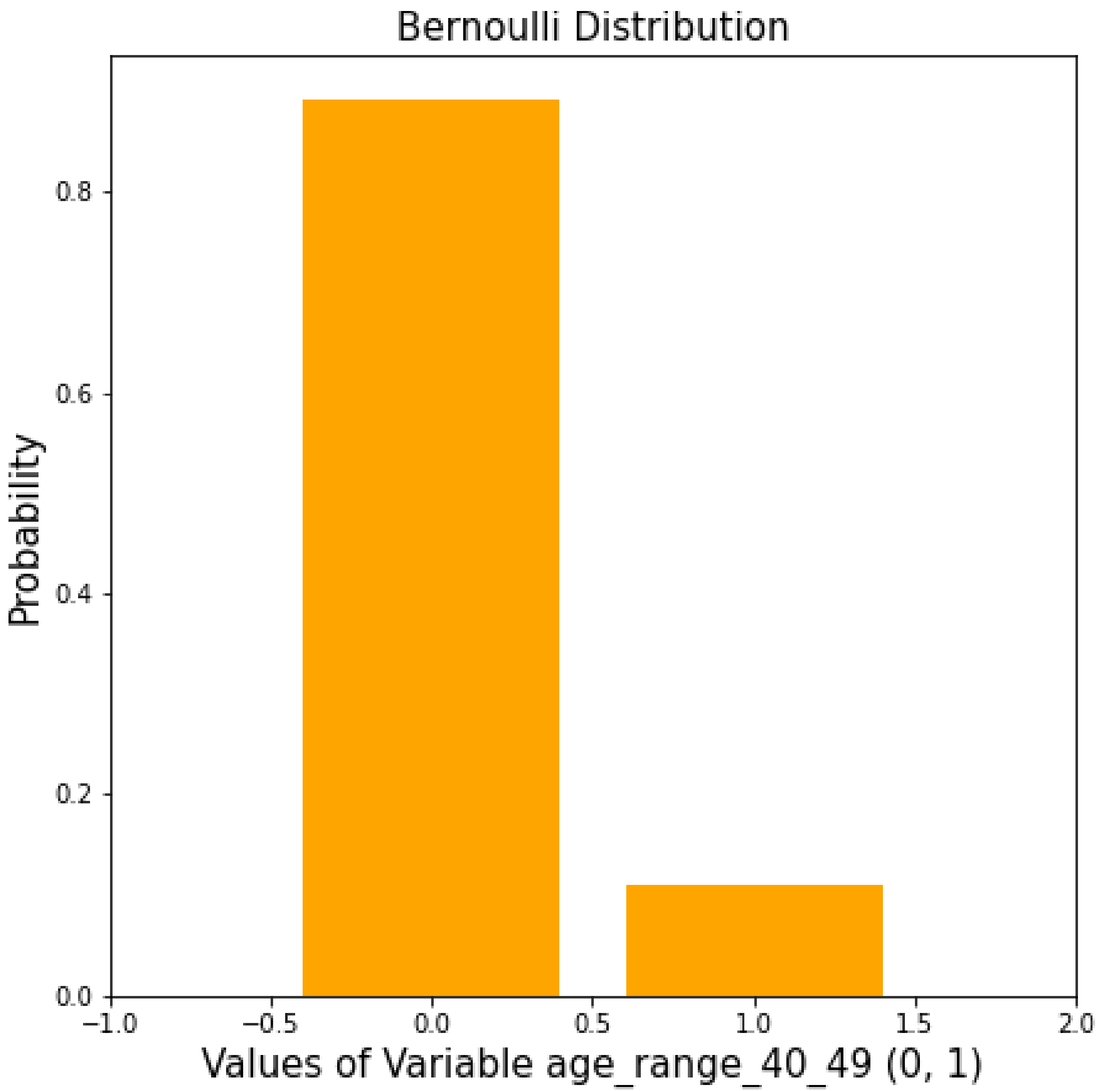

Bernoulli Distribution
Probability
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable lease_length_3 (0, 1)

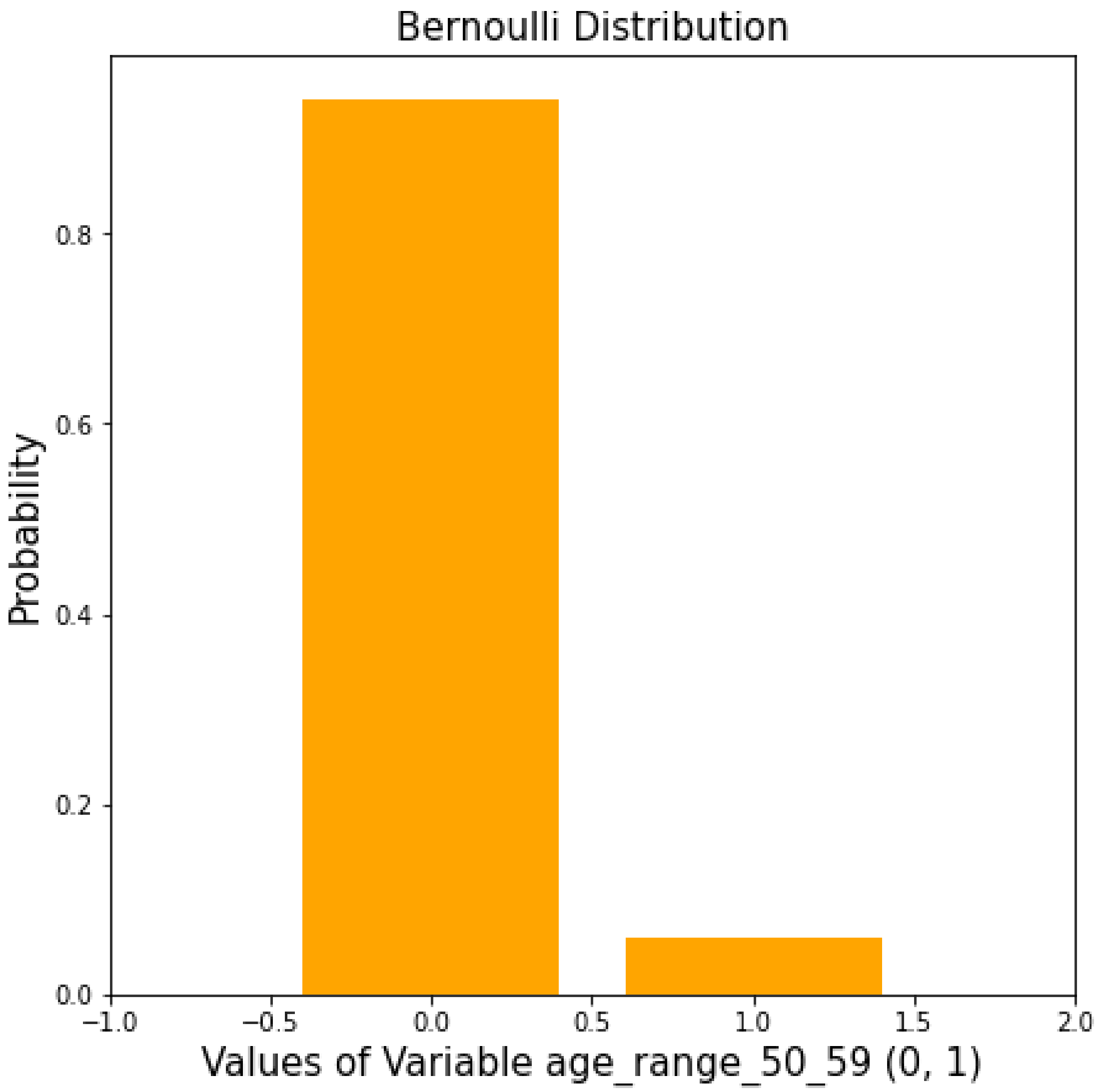

Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable lease_length_1 (0, 1)

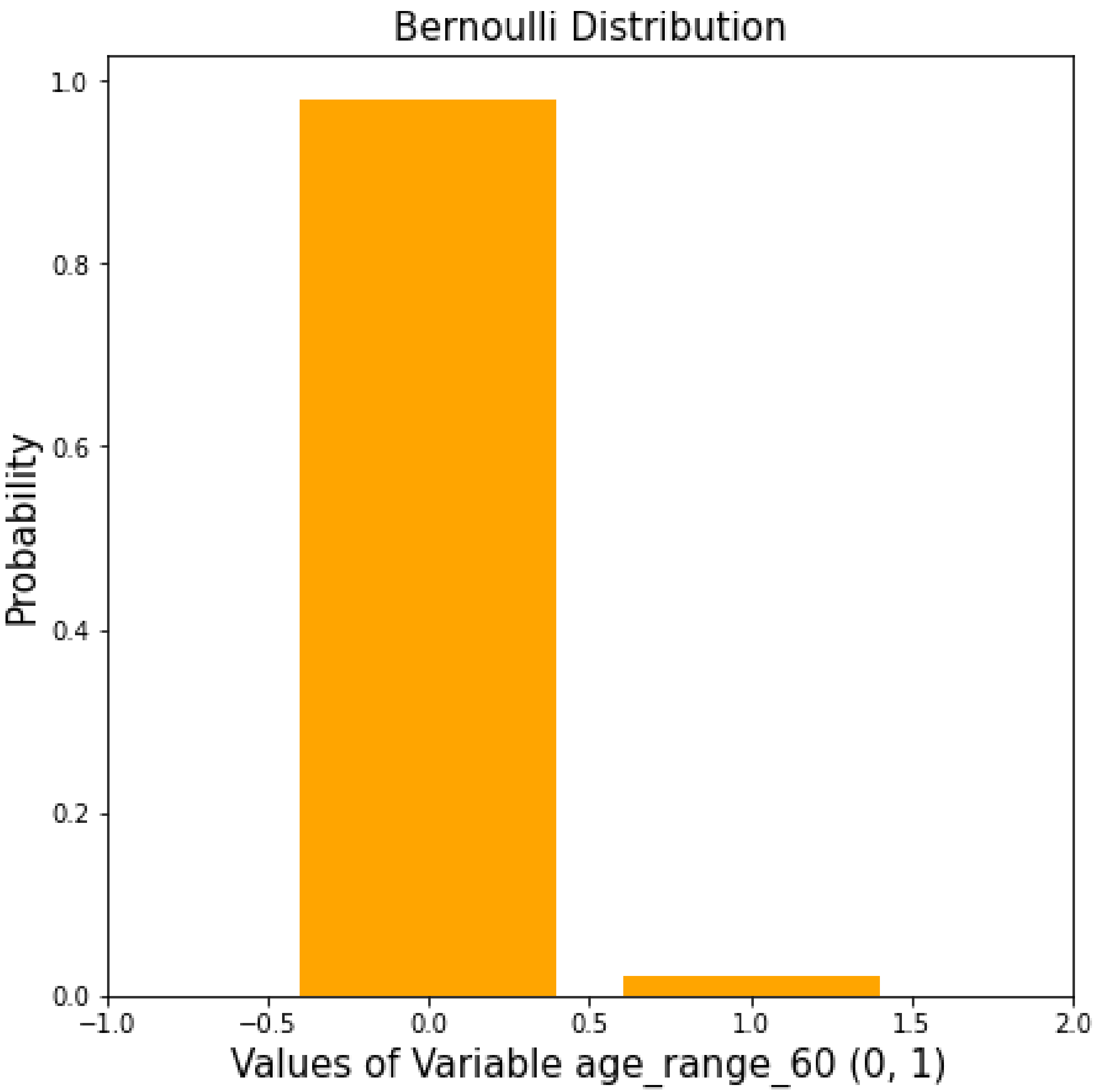

Bernoulli Distribution
Probability
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_under_24 (0, 1)

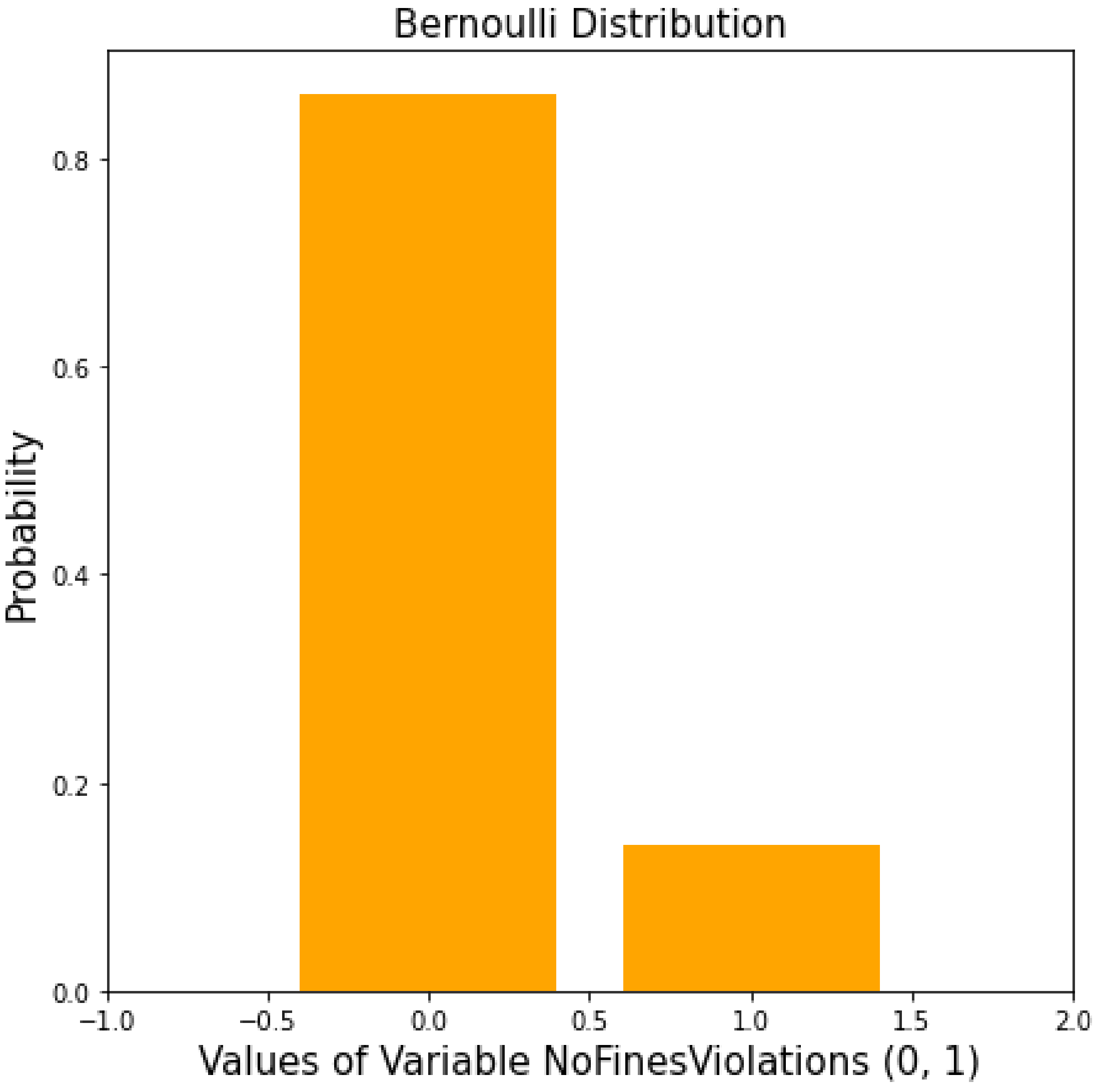

Bernoulli Distribution
Probability
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_24_29 (0, 1)

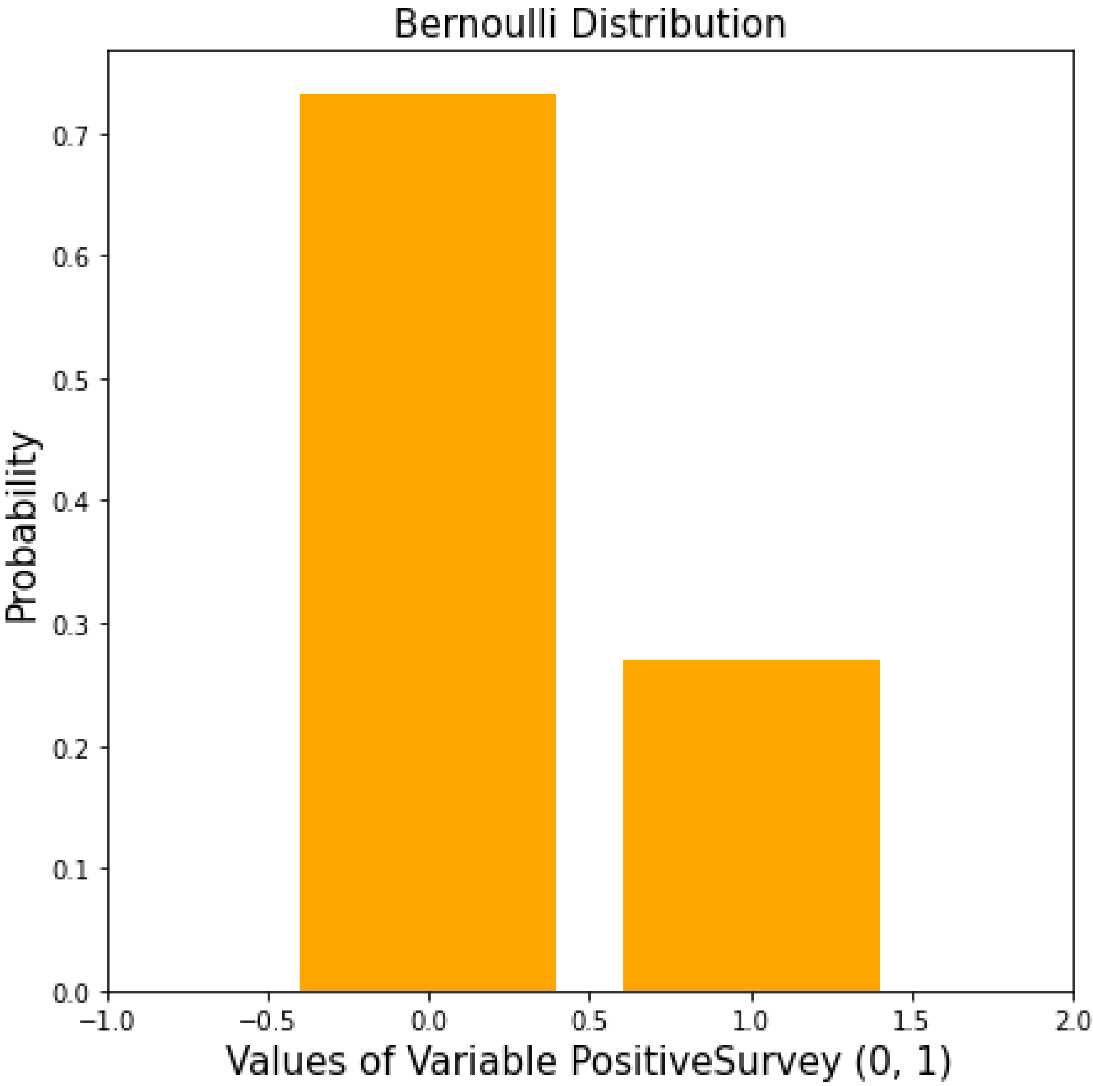

Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_30_39 (0, 1)

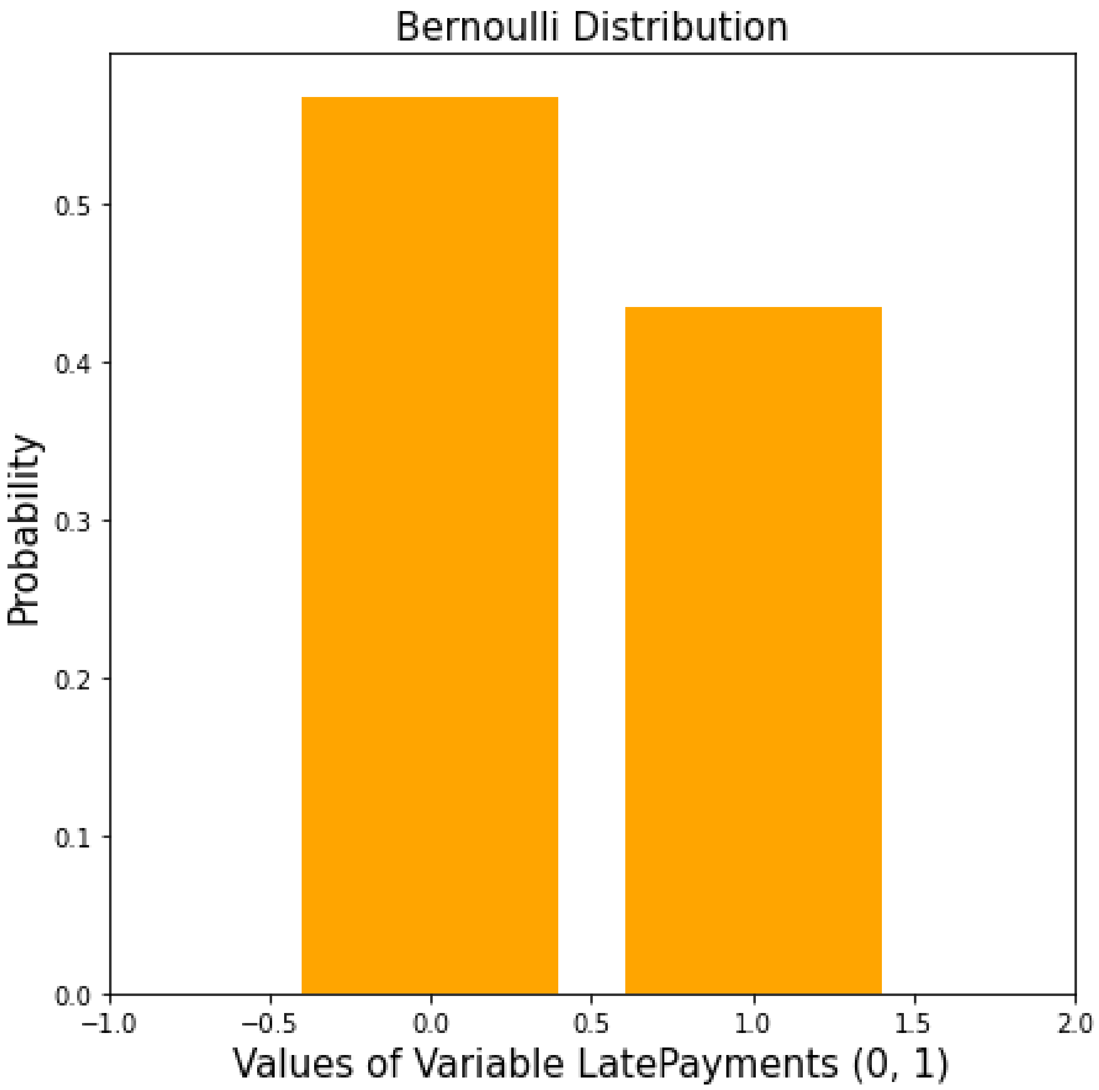

Bernoulli Distribution
Probability
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_40_49 (0, 1)

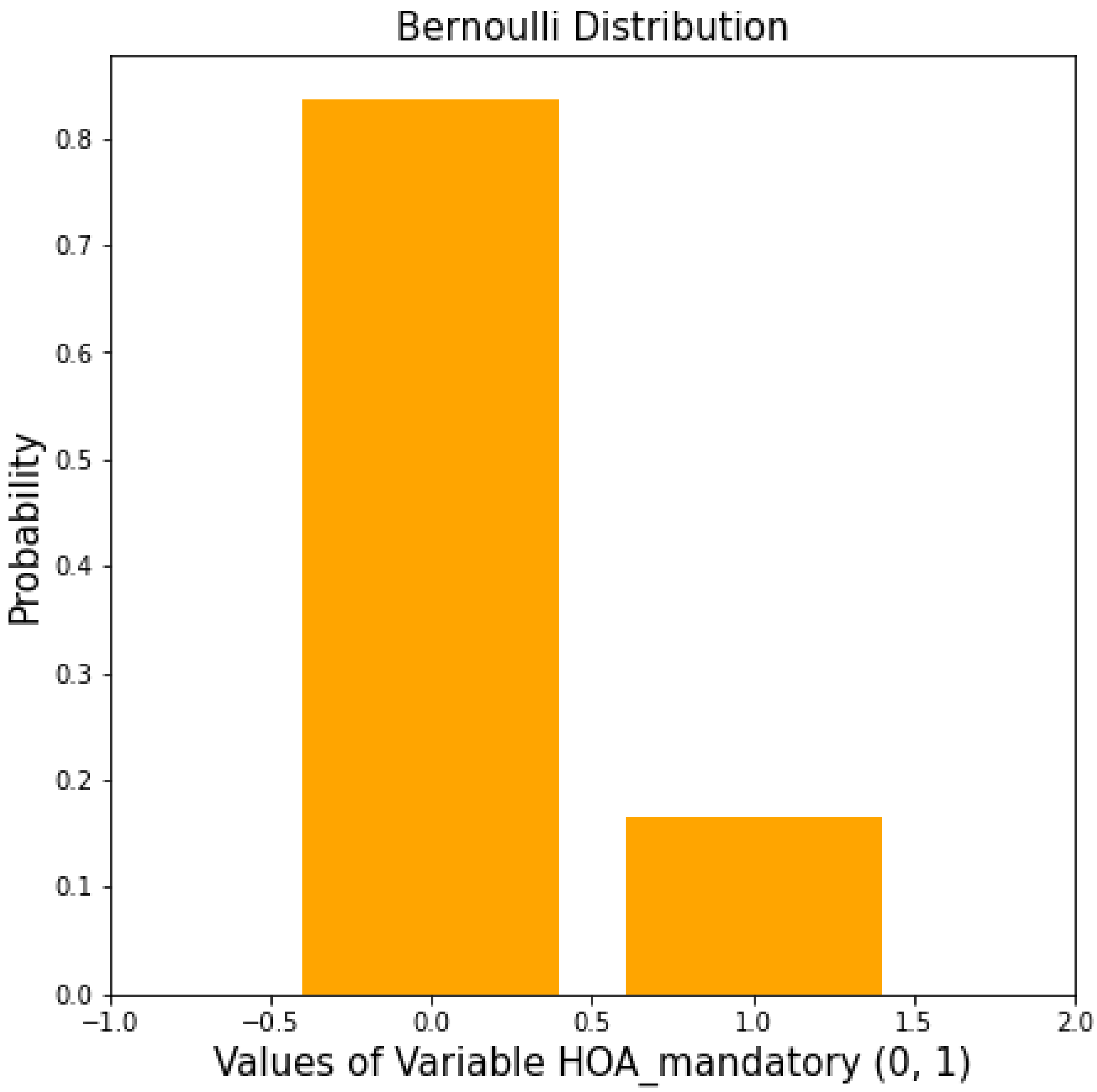

Bernoulli Distribution
Probability
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_50_59 (0, 1)

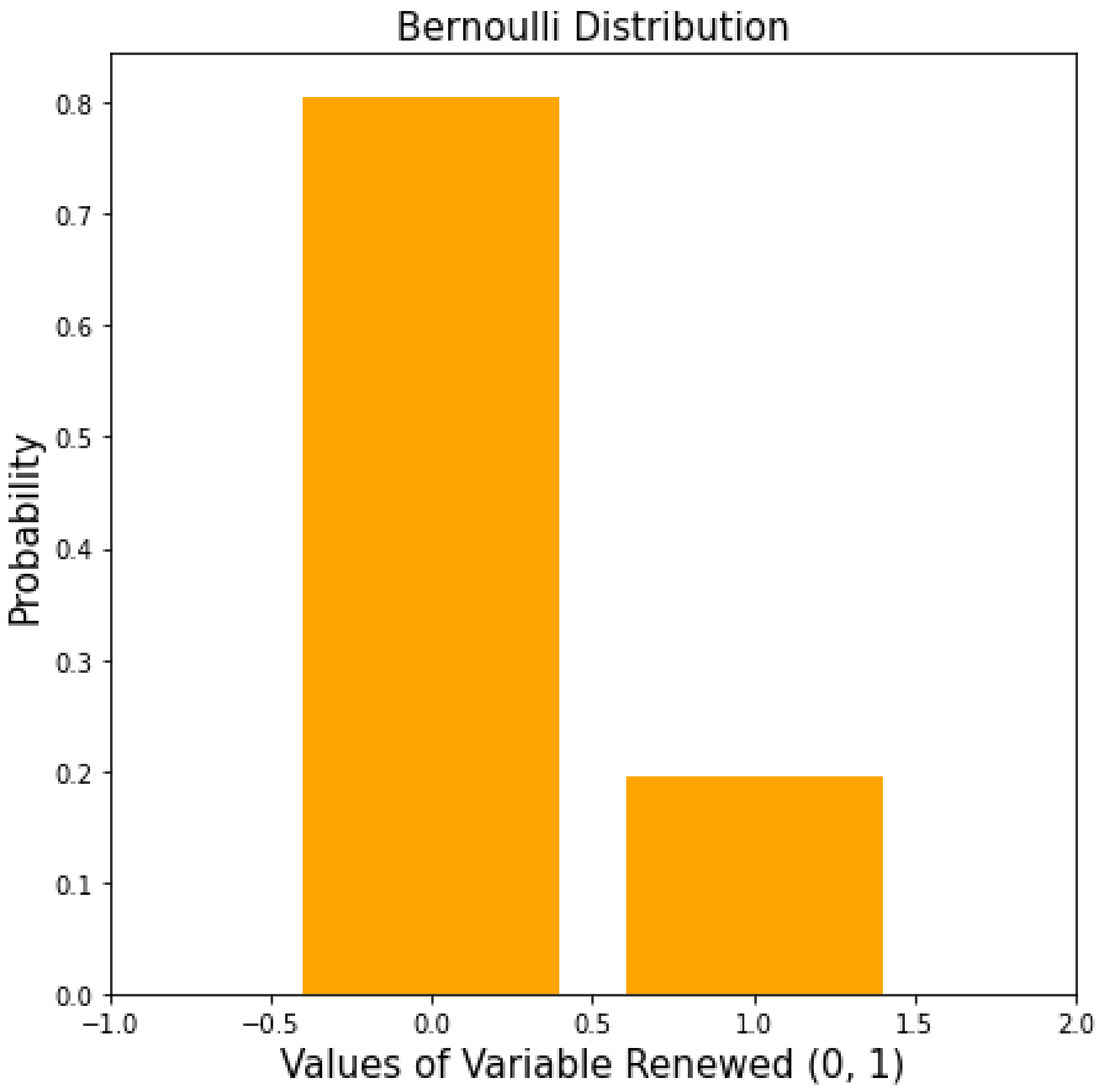

Bernoulli Distribution
Probability
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable age_range_60 (0, 1)


Bernoulli Distribution
Probability
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable NoFinesViolations (0, 1)


Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable PositiveSurvey (0, 1)


Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable LatePayments (0, 1)


Bernoulli Distribution
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Values of Variable HOA_mandatory (0, 1)



```
dataframe_renewal.iloc[:,len(dataframe_renewal.columns)-
1].value_counts()
```

```
0    64208
1    15642
Name: Renewed, dtype: int64
```

19.58 pepercentage of total residents have renewed their lease further.

Using select k best features for feature selection

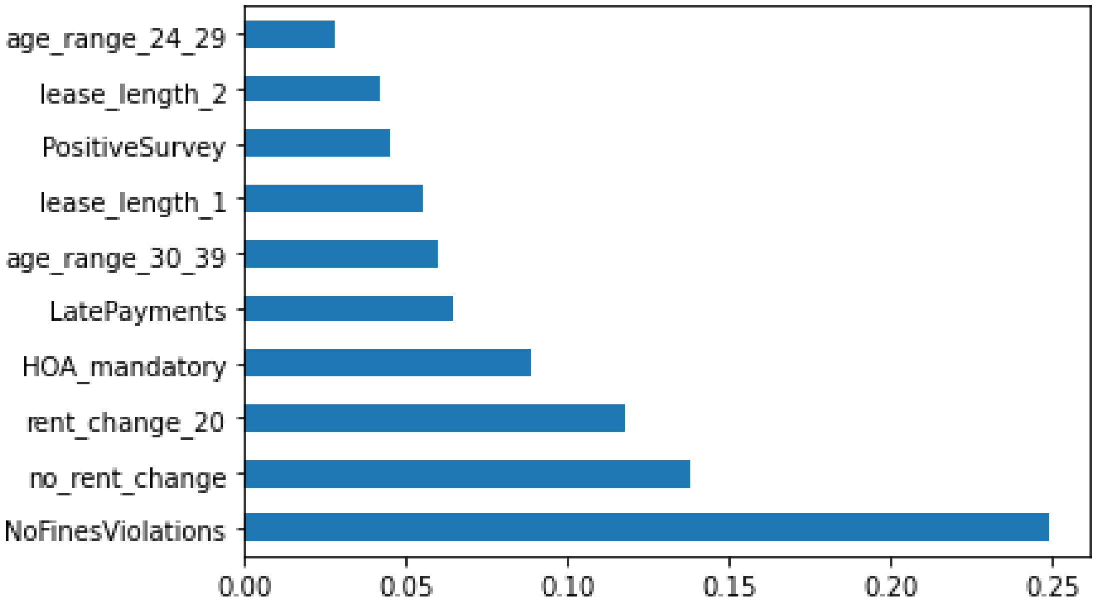
```
X = dataframe_renewal.iloc[:,1:17]
y = dataframe_renewal.iloc[:,-1]
bestfeatures = SelectKBest(score_func=chi2, k=10)
```

```
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']
print(featureScores.nlargest(10,'Score'))
```

```
                 Specs        Score
0        no_rent_change  1171.663873
12   NoFinesViolations  1141.363078
2        rent_change_20   546.303728
15        HOA_mandatory   499.392675
3        lease_length_2   443.921557
5        lease_length_1   305.374923
8       age_range_30_39   289.719040
6    age_range_under_24   203.623655
14         LatePayments   129.735829
9       age_range_40_49    48.059799
```

```
model = ExtraTreesClassifier()
model.fit(X,y)
print(model.feature_importances_)
feat_importances = pd.Series(model.feature_importances_,
index=X.columns)
feat_importances.nlargest(10).plot(kind='barh')
plt.show()
```

```
[0.13798823 0.01716024 0.11788869 0.04157493 0.01907466 0.05523355
 0.01978089 0.02763472 0.05943724 0.02532366 0.01872536 0.01239429
 0.24916766 0.04495751 0.06458268 0.08907567]
```



Observation: No violations, no_rent_change, rent_change_20, LatePayments were the highest impacting independent features

VarianceThreshold methods for feature selection

Setting the variance 10 percent. So features with less than 0.1 will have a value false.

```
from sklearn.feature_selection import VarianceThreshold
var_thres=VarianceThreshold(threshold=0.1)
var_thres.fit(dataframe_renewal.iloc[:,1:])
```

```
VarianceThreshold(threshold=0.1)
```

```
var_thres.get_support()
```

```
array([ True, False,  True,  True, False,  True, False, False,  True,
       False, False, False,  True,  True,  True,  True,  True])
```

```
dataframe_renewal.iloc[:,1:].columns
```

```
Index(['no_rent_change', 'rent_change_10', 'rent_change_20',
'lease_length_2',
       'lease_length_3', 'lease_length_1', 'age_range_under_24',
       'age_range_24_29', 'age_range_30_39', 'age_range_40_49',
       'age_range_50_59', 'age_range_60', 'NoFinesViolations',
       'PositiveSurvey', 'LatePayments', 'HOA_mandatory', 'Renewed'],
      dtype='object')
```

Observations

Below features have low variance and these features do not exist in model feature_importances.

This indicates that these

features have less impact on the dependent variable

```
dataframe_renewal['rent_change_10'].value_counts()
```

```
0    77971
1     1879
Name: rent_change_10, dtype: int64
```

```
dataframe_renewal['lease_length_3'].value_counts()
```

```
0    75259
1     4591
Name: lease_length_3, dtype: int64
```

```
dataframe_renewal['age_range_under_24'].value_counts()
```

```
0    76791
1     3059
Name: age_range_under_24, dtype: int64
```

```
dataframe_renewal['age_range_40_49'].value_counts()
```

```
0    71178
1     8672
Name: age_range_40_49, dtype: int64
```

```
dataframe_renewal['age_range_50_59'].value_counts()
```

```
0    75128
1     4722
Name: age_range_50_59, dtype: int64
```

```
dataframe_renewal['age_range_60'].value_counts()
```

```
0    78175
1     1675
Name: age_range_60, dtype: int64
```

Feature Selection-Information gain - mutual information

```
# Train test split to avoid overfitting
X_train,X_test,y_train,y_test=train_test_split(dataframe_renewal.drop(
labels=['Renewed','lease_id'], axis=1),

dataframe_renewal['Renewed'],test_size=0.3,random_state=100)
```

```
X_train.head()
```

```
       no_rent_change  rent_change_10  rent_change_20  lease_length_2
\
56768               0               1               0               0

17813               0               0               0               0

31528               1               0               0               1

30122               0               0               1               0

11090               0               0               1               0
```

```
       lease_length_3  lease_length_1  age_range_under_24
age_range_24_29  \
56768               1               0                   0
0
17813               0               0                   0
0
31528               0               0                   1
0
30122               0               1                   0
1
11090               0               1                   0
0

       age_range_30_39  age_range_40_49  age_range_50_59  age_range_60
\
56768                0                0                0             0

17813                0                0                0             0

31528                0                0                0             0

30122                0                0                0             0

11090                0                0                0             0


       NoFinesViolations  PositiveSurvey  LatePayments  HOA_mandatory

56768                  1               1             1              1

17813                  0               0             0              0

31528                  0               0             0              1

30122                  0               1             0              0

11090                  0               1             0              0
```

Determine the mutual information

```
mutual_info = mutual_info_classif(X_train, y_train)
mutual_info
```

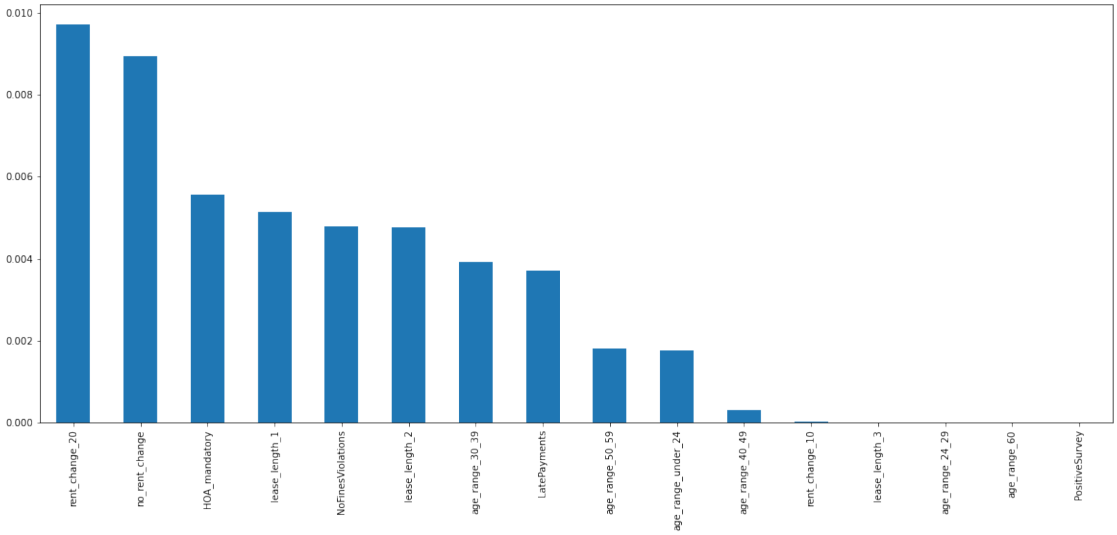
```
array([9.84239784e-03, 3.23077301e-03, 1.13970946e-02, 6.10303841e-03,
       0.00000000e+00, 6.56789683e-03, 3.78394536e-04, 6.59842213e-04,
       1.43756562e-03, 8.74230344e-05, 0.00000000e+00, 0.00000000e+00,
       8.51132478e-03, 3.01542900e-03, 5.20284414e-03, 5.63194753e-
03])
```

```
mutual_info = pd.Series(mutual_info)
mutual_info.index = X_train.columns
mutual_info.sort_values(ascending=False)
```

```
rent_change_20         0.009722
no_rent_change         0.008932
HOA_mandatory          0.005558
lease_length_1         0.005143
NoFinesViolations      0.004779
lease_length_2         0.004757
age_range_30_39        0.003907
LatePayments           0.003695
age_range_50_59        0.001798
age_range_under_24     0.001766
age_range_40_49        0.000301
rent_change_10         0.000017
lease_length_3         0.000000
age_range_24_29        0.000000
age_range_60           0.000000
PositiveSurvey         0.000000
dtype: float64
```

```
#let's plot the ordered mutual_info values per feature
mutual_info.sort_values(ascending=False).plot.bar(figsize=(20, 8))
```

```
<AxesSubplot:>
```



Observations:

rent_change_20,no_rent_change,NoFinesViolations,lease_length_1,HOA_mandatory are the top features making an highest impact

Chisquare Test For Feature Selection

```
f_p_values=chi2(X_train,y_train)
```

```
f_p_values
```

```
(array([780.14183501,  26.56792927, 379.4836922 , 318.49720696,
         11.89742588, 226.91716465, 140.62197392,   1.81715198,
        182.77954308,  38.86398527,   6.5276809 ,  40.63339053,
        810.65116529,  16.17612522,  88.28999108, 338.3505866 ]),
 array([1.12111804e-171, 2.54432455e-007, 1.61201754e-084,
3.07782890e-071,
        5.62115719e-004, 2.80334734e-051, 1.94627633e-032,
1.77652658e-001,
        1.19828020e-041, 4.54386850e-010, 1.06208343e-002,
1.83641596e-010,
        2.60810485e-178, 5.77169889e-005, 5.65287623e-021,
1.45910325e-075]))
```

```
p_values=pd.Series(f_p_values[1])
p_values.index=X_train.columns
p_values
```

```
no_rent_change        1.121118e-171
rent_change_10         2.544325e-07
rent_change_20         1.612018e-84
lease_length_2         3.077829e-71
lease_length_3         5.621157e-04
lease_length_1         2.803347e-51
age_range_under_24     1.946276e-32
age_range_24_29        1.776527e-01
age_range_30_39        1.198280e-41
age_range_40_49        4.543868e-10
age_range_50_59        1.062083e-02
age_range_60           1.836416e-10
NoFinesViolations     2.608105e-178
PositiveSurvey         5.771699e-05
LatePayments           5.652876e-21
HOA_mandatory          1.459103e-75
dtype: float64
```

```
p_values.sort_index(ascending=False)
```

```
rent_change_20         1.612018e-84
rent_change_10         2.544325e-07
no_rent_change        1.121118e-171
lease_length_3         5.621157e-04
lease_length_2         3.077829e-71
lease_length_1         2.803347e-51
age_range_under_24     1.946276e-32
age_range_60           1.836416e-10
age_range_50_59        1.062083e-02
age_range_40_49        4.543868e-10
age_range_30_39        1.198280e-41
age_range_24_29        1.776527e-01
PositiveSurvey         5.771699e-05
NoFinesViolations     2.608105e-178
LatePayments           5.652876e-21
HOA_mandatory          1.459103e-75
dtype: float64
```

Observation

NoFinesViolations, no_rent_change, rent_change_20, HOA_mandatory, lease_length_2 are the top features having the hightest impact

By looking at all the above steps we see that no rent change, no fines violated, Hoa mandatory, rentchange20 are the features having the most impact on dependent feature.

Calculating the percentage impact on the renewed feature by no_rent_change

```
listval_no_rent_change=[]
for itr in range(len(dataframe_renewal.index)):
    if dataframe_renewal.Renewed[itr]==1:

listval_no_rent_change.append(dataframe_renewal.no_rent_change[itr])
print(listval_no_rent_change.count(1))
print(listval_no_rent_change.count(0))
```

```
5270
10372
```

Observation:

From no_rent_change around 34 percent of the residents renew if no_rent_change = 1 that means most of the people who renewed has their rent increased

Below clearly indicates that there are no records with no_rent_change = 1 and rent_change =1 which indicates: if there is no change in rent then there is no change in the rent as well

```
check=[]
for itr in range(len(dataframe_renewal.index)):
    if (dataframe_renewal.no_rent_change[itr]==1 and
(dataframe_renewal.rent_change_10[itr]==1 or

dataframe_renewal.rent_change_20[itr]==1)) :
        check.append(dataframe_renewal.no_rent_change[itr])
print(len(check))
```

```
0
```

```
check=[]
for itr in range(len(dataframe_renewal.index)):
    if (dataframe_renewal.no_rent_change[itr]==0 and
(dataframe_renewal.rent_change_10[itr]==0 and

dataframe_renewal.rent_change_20[itr]==0)) :
        check.append(dataframe_renewal.no_rent_change[itr])
print(len(check))
```

```
13828
```

13828 records exists that indicates: There was a change in the rent and neither it is increased by 10 nor 20 percent.

Calculating the percentage impact on the renewed feature by rent_change_20

```
listval_rent_change_20=[]
for itr in range(len(dataframe_renewal.index)):
    if dataframe_renewal.Renewed[itr]==1:

listval_rent_change_20.append(dataframe_renewal.rent_change_20[itr])
```

```
print(listval_rent_change_20.count(1))
print(listval_rent_change_20.count(0))
```

```
7102
8540
```

From rent_change_20 around 55 percent of the residents renew if rent_change_20 = 0

that means people prefer having no change is rent although not a high percentage

Calculating the percentage impact on the renewed feature by NoFinesViolations

```
listval_NoFinesViolations=[]
for itr in range(len(dataframe_renewal.index)):
    if dataframe_renewal.Renewed[itr]==1:

listval_NoFinesViolations.append(dataframe_renewal.NoFinesViolations[i
tr])
print(listval_NoFinesViolations.count(1))
print(listval_NoFinesViolations.count(0))
```

```
3592
12050
```

From NoFinesViolations 77 percentage of the residents renew if NoFinesViolations = 0

that means most of the people who renewed the lease have violations

Calculating the percentage impact on the renewed feature by Hoa mandatory

```
listval_HOA_mandatory=[]
for itr in range(len(dataframe_renewal.index)):
    if dataframe_renewal.Renewed[itr]==1:

listval_HOA_mandatory.append(dataframe_renewal.HOA_mandatory[itr])
```

```
print(listval_HOA_mandatory.count(1))
print(listval_HOA_mandatory.count(0))
```

```
1552
14090
```

From HOA_mandatory around 90 percentage of residents renew if HOA_mandatory = 0

that means highest people prefer having No mandatory fees on the lease