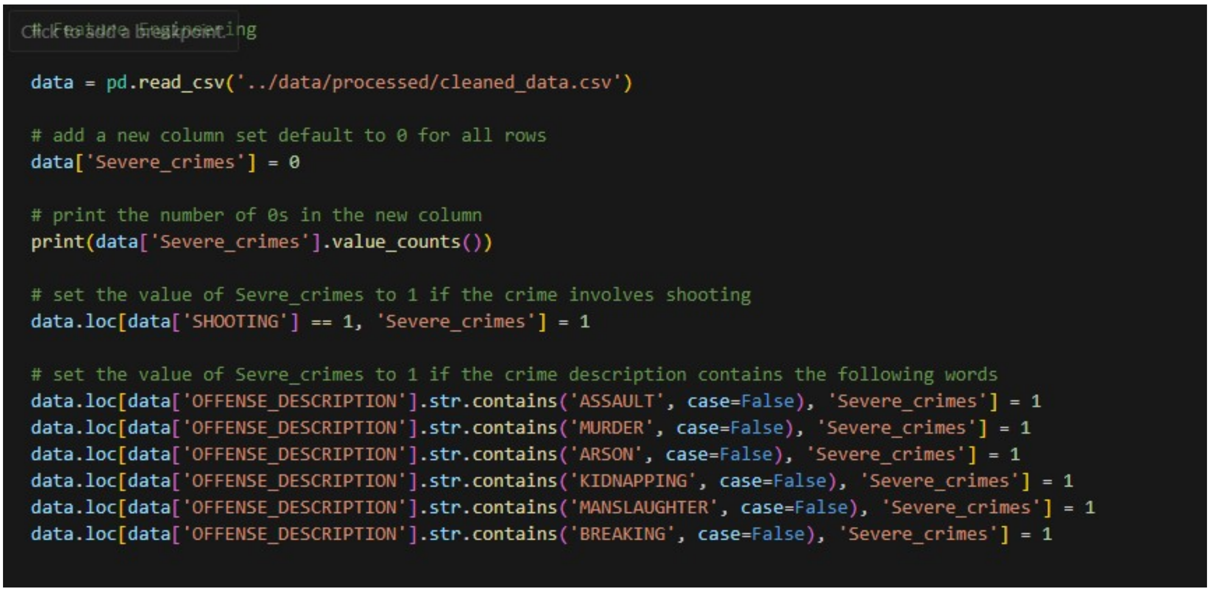
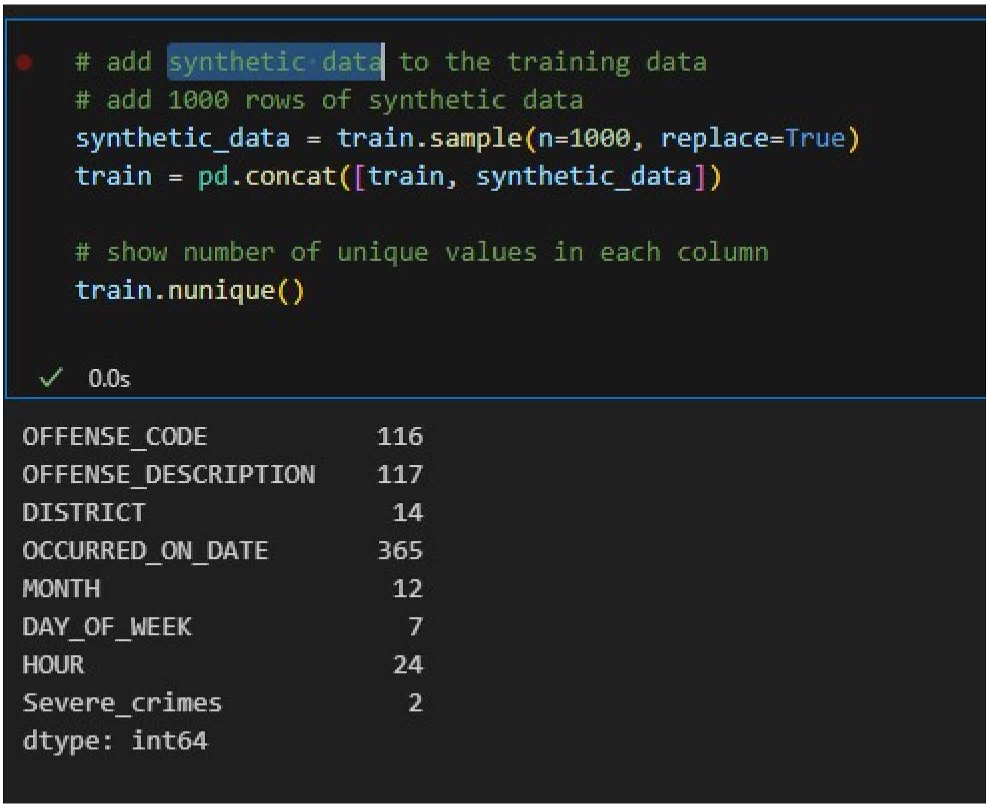
The column Severe_Crime is the result of feature engineering, In which,



```
# Feature Engineering

data = pd.read_csv('../data/processed/cleaned_data.csv')

# add a new column set default to 0 for all rows
data['Severe_crimes'] = 0

# print the number of 0s in the new column
print(data['Severe_crimes'].value_counts())

# set the value of Sevre_crimes to 1 if the crime involves shooting
data.loc[data['SHOOTING'] == 1, 'Severe_crimes'] = 1

# set the value of Sevre_crimes to 1 if the crime description contains the following words
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('ASSAULT', case=False), 'Severe_crimes'] = 1
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('MURDER', case=False), 'Severe_crimes'] = 1
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('ARSON', case=False), 'Severe_crimes'] = 1
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('KIDNAPPING', case=False), 'Severe_crimes'] = 1
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('MANSLAUGHTER', case=False), 'Severe_crimes'] = 1
data.loc[data['OFFENSE_DESCRIPTION'].str.contains('BREAKING', case=False), 'Severe_crimes'] = 1
```

This creates a new column that only have categorical values of 1 and 0. And this indicates that if this crimes needs more law enforcement force and more equipment to handle.

For synthetic data, 1000 synthetic data was added for testing purpose.



```
# add synthetic data to the training data
# add 1000 rows of synthetic data
synthetic_data = train.sample(n=1000, replace=True)
train = pd.concat([train, synthetic_data])

# show number of unique values in each column
train.nunique()
```

```
OFFENSE_CODE            116
OFFENSE_DESCRIPTION     117
DISTRICT                 14
OCCURRED_ON_DATE        365
MONTH                    12
DAY_OF_WEEK               7
HOUR                     24
Severe_crimes             2
dtype: int64
```

If adding more data will lead to a better result of the model accuracy, larger number of synthetic will be used. If this shows no help, the original dataset will be used for final model training.