```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
import pandas as pd
df = pd.read_csv('fraudTest.csv')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19466 entries, 0 to 19465
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Unnamed: 0             19466 non-null  int64  
 1   trans_date_trans_time  19466 non-null  object 
 2   cc_num                 19466 non-null  int64  
 3   merchant               19466 non-null  object 
 4   category               19466 non-null  object 
 5   amt                    19466 non-null  float64
 6   first                  19466 non-null  object 
 7   last                   19466 non-null  object 
 8   gender                 19466 non-null  object 
 9   street                 19466 non-null  object 
 10  city                   19465 non-null  object 
 11  state                  19465 non-null  object 
 12  zip                    19465 non-null  float64
 13  lat                    19465 non-null  float64
 14  long                   19465 non-null  float64
 15  city_pop               19465 non-null  float64
 16  job                    19465 non-null  object 
 17  dob                    19465 non-null  object 
 18  trans_num              19465 non-null  object 
 19  unix_time              19465 non-null  float64
 20  merch_lat              19465 non-null  float64
 21  merch_long             19465 non-null  float64
 22  is_fraud               19465 non-null  float64
dtypes: float64(9), int64(2), object(12)
memory usage: 3.4+ MB
```

```
df.isnull().sum()
```

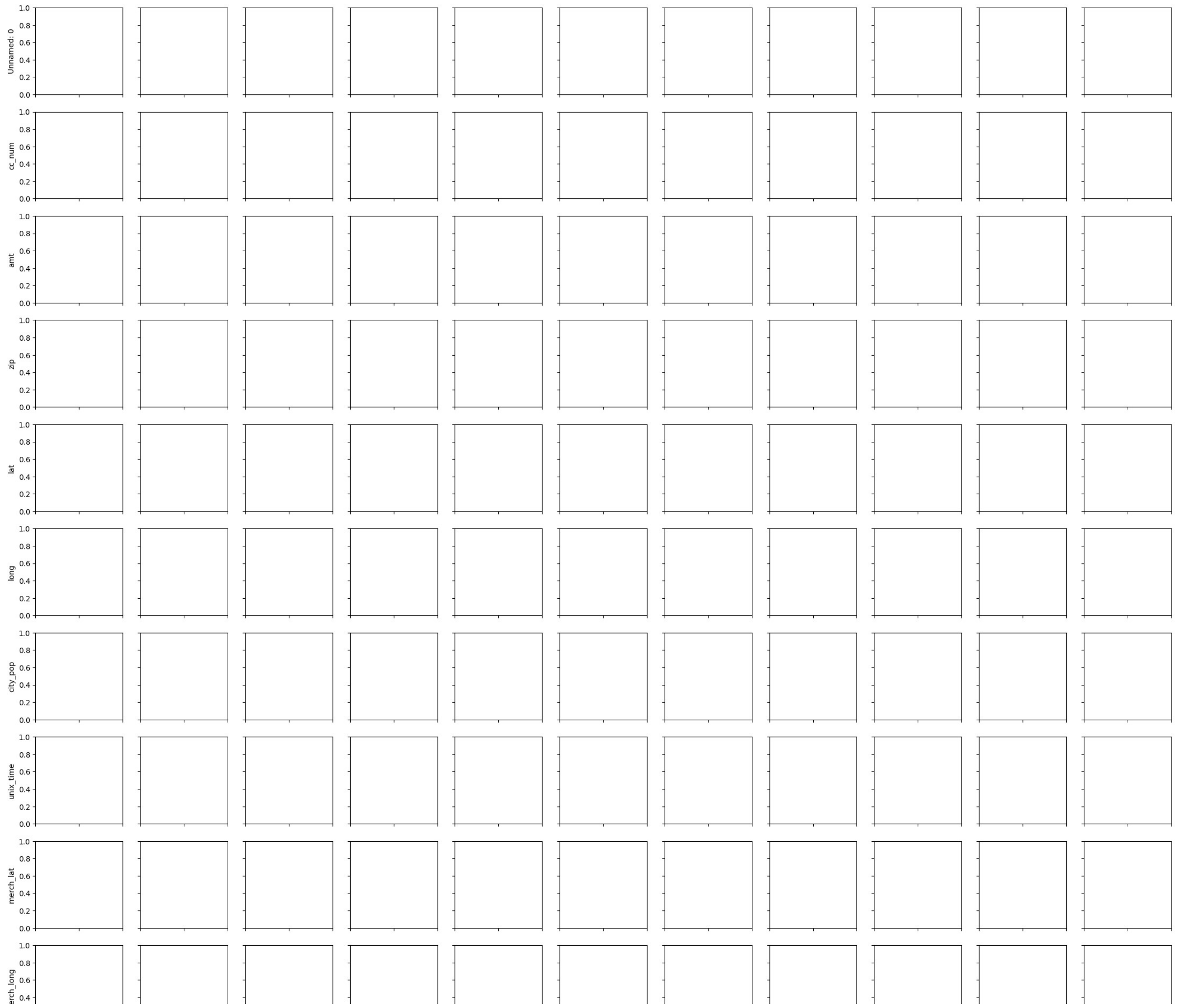
	0
Unnamed: 0	0

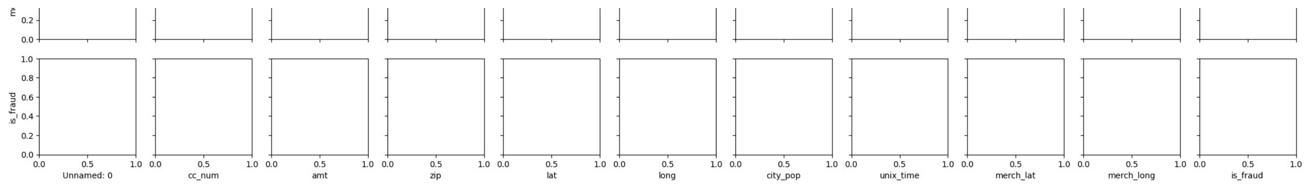
trans_date_trans_time	0
cc_num	0
merchant	0
category	0
amt	0
first	0
last	0
gender	0
street	0
city	1
state	1
zip	1
lat	1
long	1
city_pop	1
job	1
dob	1
trans_num	1
unix_time	1
merch_lat	1
merch_long	1

```
import seaborn as sns
sns.pairplot(df,hue='Class',palette='Set1')
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in get_loc(self,
key)
   3804         try:
-> 3805             return self._engine.get_loc(casted_key)
   3806         except KeyError as err:

index.pyx in pandas._libs.index.IndexEngine.get_loc()

index.pyx in pandas._libs.index.IndexEngine.get_loc()
```

```
pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in
pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'Class'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
                                   4 frames
/usr/local/lib/python3.11/dist-packages/pandas/core/indexes/base.py in get_loc(self,
key)
   3810                 ):
   3811                     raise InvalidIndexError(key)
-> 3812                 raise KeyError(key) from err
   3813             except TypeError:
   3814                 # If we have a listlike key, _check_indexing_error will raise

KeyError: 'Class'
```





Next steps: Explain error

```
df.head()
```

	Unnamed: 0	trans_date_trans_time	cc_num	merchant	category
0	0	2020-06-21 12:14:25	2291163933867244	fraud_Kirlin and Sons	personal_care
1	1	2020-06-21 12:14:33	3573030041201292	fraud_Sporer-Keebler	personal_care
2	2	2020-06-21 12:14:53	3598215285024754	fraud_Swaniawski, Nitzsche and Welch	health_fitness
3	3	2020-06-21 12:15:15	3591919803438423	fraud_Haley Group	misc_pos
4	4	2020-06-21 12:15:17	3526826139003047	fraud_Johnston-Casper	travel

5 rows × 23 columns

```
from sklearn.model_selection import train_test_split

X = df.drop('is_fraud',axis=1)
y = df['is_fraud']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_
```

```
from sklearn.tree import DecisionTreeClassifier

dtree = DecisionTreeClassifier(criterion='entropy', random_sta

dtree.fit(X_train,y_train)
```

Show hidden output

Next steps: Explain error

Start coding or generate with AI.

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
import pandas as pd

# Assuming 'df' is your DataFrame

# Convert the 'trans_date_trans_time' column to datetime objec
df['trans_date_trans_time'] = pd.to_datetime(df['trans_date_tr

# Extract features from the datetime column
df['transaction_hour'] = df['trans_date_trans_time'].dt.hour
df['transaction_day'] = df['trans_date_trans_time'].dt.day
df['transaction_month'] = df['trans_date_trans_time'].dt.month
df['transaction_year'] = df['trans_date_trans_time'].dt.year

# Convert 'misc_net' and 'misc_pos' to numerical representatio
# Convert 'misc_net' and 'misc_pos' to numerical representatio
# Assuming 'misc_net' and 'misc_pos' contain values like 'misc_
# Replace with 1 if the value is present, 0 otherwise
for col in ['misc_net', 'misc_pos']:
    df[col] = df[col].apply(lambda x: 1 if isinstance(x, str) 
    #This line checks if the value in the column is a string. 
    #it assumes it represents a categorical value (like 'misc_
    # If it's not a string, it assumes it's already a numerica


# Drop the original 'trans_date_trans_time' column
# Also drop other irrelevant columns such as 'cc_num', 'mercha
# If there are more string features replace them with numerica
X = df.drop(['is_fraud', 'trans_date_trans_time', 'cc_num', 'm
y = df['is_fraud']


X_train, X_test, y_train, y_test = train_test_split(X, y, test_

dtree = DecisionTreeClassifier(criterion='entropy', random_sta

dtree.fit(X_train,y_train) # Now the fitting should work witho

df[col] = df[col].apply(lambda x: 1 if isinstance(x, str) else

predictions = dtree.predict(X_test)
predictions



from sklearn.metrics import classification_report,confusion_ma
print(classification_report(y_test,predictions))
```