# Initial_Code_and_Analysis

June 27, 2023

### Credit Card Fraud Transaction Data

Shenoy, K. (2019). Credit Card Transactions Fraud Detection Dataset. Kaggle.com. https://www.kaggle.com/datasets/kartik2112/fraud-detection

***Reading in the dataset and importing relevant packages***

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import datetime
import calendar
from sklearn.preprocessing import RobustScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import RandomOverSampler
from sklearn.linear_model import LogisticRegression
from sklearn import tree
from sklearn.neighbors import KNeighborsClassifier
from sklearn.pipeline import Pipeline
from imblearn.pipeline import Pipeline as imbpipeline
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import make_scorer, accuracy_score, precision_score,␣
↪recall_score, f1_score, roc_auc_score
```

[2]:
```
fraud_train = pd.read_csv('/content/drive/MyDrive/CIND 820/fraudTrain.csv')
fraud_train.shape[0]
```

[2]: 1296675

[3]:
```
fraud_test = pd.read_csv('/content/drive/MyDrive/CIND 820/fraudTest.csv')
fraud_test.shape[0]
```

[3]: 555719

[4]:
```python
fraud = pd.concat([fraud_train, fraud_test])
fraud.shape[0] # Gives total number of observations
```

[4]: 1852394

[ ]:
```python
fraud.head(5)
```

[ ]:
```
   Unnamed: 0 trans_date_trans_time            cc_num  \
0           0   2019-01-01 00:00:18  2703186189652095
1           1   2019-01-01 00:00:44      630423337322
2           2   2019-01-01 00:00:51    38859492057661
3           3   2019-01-01 00:01:16  3534093764340240
4           4   2019-01-01 00:03:06   375534208663984

                             merchant       category     amt      first  \
0          fraud_Rippin, Kub and Mann       misc_net    4.97   Jennifer
1     fraud_Heller, Gutmann and Zieme    grocery_pos  107.23  Stephanie
2                fraud_Lind-Buckridge  entertainment  220.11     Edward
3  fraud_Kutch, Hermiston and Farrell  gas_transport   45.00     Jeremy
4                 fraud_Keeling-Crist       misc_pos   41.96      Tyler

      last gender                        street  …      lat      long  \
0    Banks      F                561 Perry Cove  …  36.0788  -81.1781
1     Gill      F  43039 Riley Greens Suite 393  …  48.8878 -118.2105
2  Sanchez      M      594 White Dale Suite 530  …  42.1808 -112.2620
3    White      M   9443 Cynthia Court Apt. 038  …  46.2306 -112.1138
4   Garcia      M              408 Bradley Rest  …  38.4207  -79.4629

   city_pop                                job         dob  \
0      3495          Psychologist, counselling  1988-03-09
1       149  Special educational needs teacher  1978-06-21
2      4154        Nature conservation officer  1962-01-19
3      1939                    Patent attorney  1967-01-12
4        99     Dance movement psychotherapist  1986-03-28

                          trans_num   unix_time  merch_lat  merch_long  \
0  0b242abb623afc578575680df30655b9  1325376018  36.011293  -82.048315
1  1f76529f8574734946361c461b024d99  1325376044  49.159047 -118.186462
2  a1a22d70485983eac12b5b88dad1cf95  1325376051  43.150704 -112.154481
3  6b849c168bdad6f867558c3793159a81  1325376076  47.034331 -112.561071
4  a41d7549acf90789359a9aa5346dcb46  1325376186  38.674999  -78.632459

   is_fraud
0         0
1         0
2         0
3         0
```

```
4        0

[5 rows x 23 columns]
```

***Data Cleaning***

```
[ ]: fraud.dtypes
```

```
[ ]: Unnamed: 0                 int64
     trans_date_trans_time     object
     cc_num                     int64
     merchant                  object
     category                  object
     amt                      float64
     first                     object
     last                      object
     gender                    object
     street                    object
     city                      object
     state                     object
     zip                        int64
     lat                      float64
     long                     float64
     city_pop                   int64
     job                       object
     dob                       object
     trans_num                 object
     unix_time                  int64
     merch_lat                float64
     merch_long               float64
     is_fraud                   int64
     dtype: object
```

```
[ ]: pd.value_counts(fraud.dtypes) # Shows the frequency of the relevant data types␣
     ↪in data set
```

```
[ ]: object     12
     int64       6
     float64     5
     dtype: int64
```

Checking the data types for all the variables in the dataset, we can see that trans_date_trans_time and dob (date of birth) are of an 'object' data type, which is not correct.

Knowing this, both of these variables need to be converted into their appropriate data type, which is datetime.

[5]:
```python
fraud[['trans_date_trans_time', 'dob']] = fraud[['trans_date_trans_time',␣
↪'dob']].apply(pd.to_datetime)
fraud.dtypes[['trans_date_trans_time', 'dob']] # Check
```

[5]:
```
trans_date_trans_time    datetime64[ns]
dob                      datetime64[ns]
dtype: object
```

Looking first with trans_date_trans_time, we can extract the month, year, and day of the week of each observation and create a variable for each one.

[6]:
```python
# For transaction month
fraud['month'] = fraud['trans_date_trans_time'].dt.month_name()
fraud['month'].head(5) # Check
```

[6]:
```
0    January
1    January
2    January
3    January
4    January
Name: month, dtype: object
```

[ ]:
```python
fraud['month'].value_counts() # Displays the frequency for each month
```

[ ]:
```
December     280598
August       176118
June         173869
July         172444
May          146875
March        143789
November     143056
September    140185
October      138106
April        134970
January      104727
February      97657
Name: month, dtype: int64
```

Examing the distribution of the number of observations by month, it is shown that December appears more frequently in the data set where 280,598 observations out of 1,852,394 observations are in that month. It can be inferred that this is most likely due to people shopping for the holiday season.

[7]:
```python
# For transaction day
fraud['day'] = fraud['trans_date_trans_time'].dt.strftime('%A')
```

[ ]:
```python
fraud['day'].value_counts() # Displays the frequency of each day of the week
```

```
[ ]: Monday       369418
     Sunday       343677
     Tuesday      270340
     Saturday     263227
     Friday       215078
     Thursday     206741
     Wednesday    183913
     Name: day, dtype: int64
```

As for days, approximately 19.94% of the observations are assigned to Monday, meaning that most people conduct their transactions on a Monday.

```
[8]: # For transaction year
     fraud['year'] = fraud['trans_date_trans_time'].dt.strftime('%Y')
```

```
[ ]: fraud['year'].value_counts() # Displays the distribution between the 2 years
```

```
[ ]: 2020    927544
     2019    924850
     Name: year, dtype: int64
```

In terms of years, both 2019 and 2020 are almost equally repsented in the data set, 49.93% and 50.07% respectively.

Similarly, we can use 'dob' to create an age variable.

```
[9]: difference = fraud['trans_date_trans_time'] - fraud['dob']
     fraud['age'] = difference.dt.days // 365
```

```
[ ]: fraud['age'].head(5) # Check
```

```
[ ]: 0    30
     1    40
     2    56
     3    52
     4    32
     Name: age, dtype: int64
```

Now that we have gathered and generated a year, month, weekday, and age variable from 'trans_date_trans_time' and 'dob', we can drop both of these variables from the dataset. As well, we can also drop 'Unnamed: 0' as it does not contain valuable information to aid in our analysis.

```
[10]: fraud = fraud.drop(['trans_date_trans_time', 'dob', 'Unnamed: 0'], axis=1)
```

Looking back at the data types above, we see that 'cc_num' (credit card number) and zip' are of numeric data type (int64). Since credit card number serves as an identifer and that zip code is a type of geographic information, it would need to be converted into a categorical data type as peforming numerical calculations for both would not output anything useful.

[11]:
```python
fraud[['cc_num','zip']] = fraud[['cc_num', 'zip']].apply(str)
```

***Exploring the data***

Checking again the first 5 observations of the dataset.

[ ]:
```python
fraud.head(5)
```

[ ]:
```
                                               cc_num  \
0  0         2703186189652095\n1             6304…
1  0         2703186189652095\n1             6304…
2  0         2703186189652095\n1             6304…
3  0         2703186189652095\n1             6304…
4  0         2703186189652095\n1             6304…

                              merchant       category     amt      first  \
0           fraud_Rippin, Kub and Mann       misc_net    4.97   Jennifer
1      fraud_Heller, Gutmann and Zieme    grocery_pos  107.23  Stephanie
2                 fraud_Lind-Buckridge  entertainment  220.11     Edward
3  fraud_Kutch, Hermiston and Farrell  gas_transport   45.00     Jeremy
4                  fraud_Keeling-Crist       misc_pos   41.96      Tyler

      last gender                        street            city state  …  \
0    Banks      F                561 Perry Cove  Moravian Falls    NC  …
1     Gill      F  43039 Riley Greens Suite 393          Orient    WA  …
2  Sanchez      M      594 White Dale Suite 530      Malad City    ID  …
3    White      M   9443 Cynthia Court Apt. 038         Boulder    MT  …
4   Garcia      M              408 Bradley Rest        Doe Hill    VA  …

                                 job                         trans_num  \
0          Psychologist, counselling  0b242abb623afc578575680df30655b9
1  Special educational needs teacher  1f76529f8574734946361c461b024d99
2        Nature conservation officer  a1a22d70485983eac12b5b88dad1cf95
3                    Patent attorney  6b849c168bdad6f867558c3793159a81
4     Dance movement psychotherapist  a41d7549acf90789359a9aa5346dcb46

    unix_time  merch_lat  merch_long is_fraud    month      day  year  age
0  1325376018  36.011293  -82.048315        0  January  Tuesday  2019   30
1  1325376044  49.159047 -118.186462        0  January  Tuesday  2019   40
2  1325376051  43.150704 -112.154481        0  January  Tuesday  2019   56
3  1325376076  47.034331 -112.561071        0  January  Tuesday  2019   52
4  1325376186  38.674999  -78.632459        0  January  Tuesday  2019   32

[5 rows x 24 columns]
```

[ ]:
```python
fraud.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 1852394 entries, 0 to 555718
Data columns (total 24 columns):
 #   Column      Dtype  
---  ------      -----  
 0   cc_num      object 
 1   merchant    object 
 2   category    object 
 3   amt         float64
 4   first       object 
 5   last        object 
 6   gender      object 
 7   street      object 
 8   city        object 
 9   state       object 
 10  zip         object 
 11  lat         float64
 12  long        float64
 13  city_pop    int64  
 14  job         object 
 15  trans_num   object 
 16  unix_time   int64  
 17  merch_lat   float64
 18  merch_long  float64
 19  is_fraud    int64  
 20  month       object 
 21  day         object 
 22  year        object 
 23  age         int64  
dtypes: float64(5), int64(4), object(15)
memory usage: 353.3+ MB
```

[ ]:
```
fraud.describe()
```

[ ]:
```
                amt           lat          long      city_pop     unix_time  \
count  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06   
mean   7.006357e+01  3.853931e+01 -9.022783e+01  8.864367e+04  1.358674e+09   
std    1.592540e+02  5.071470e+00  1.374789e+01  3.014876e+05  1.819508e+07   
min    1.000000e+00  2.002710e+01 -1.656723e+02  2.300000e+01  1.325376e+09   
25%    9.640000e+00  3.466890e+01 -9.679800e+01  7.410000e+02  1.343017e+09   
50%    4.745000e+01  3.935430e+01 -8.747690e+01  2.443000e+03  1.357089e+09   
75%    8.310000e+01  4.194040e+01 -8.015800e+01  2.032800e+04  1.374581e+09   
max    2.894890e+04  6.669330e+01 -6.795030e+01  2.906700e+06  1.388534e+09   

          merch_lat    merch_long      is_fraud           age  
count  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06  
mean   3.853898e+01 -9.022794e+01  5.210015e-03  4.579690e+01  
std    5.105604e+00  1.375969e+01  7.199217e-02  1.742393e+01  
```

```
min    1.902742e+01 -1.666716e+02  0.000000e+00  1.300000e+01
25%    3.474012e+01 -9.689944e+01  0.000000e+00  3.200000e+01
50%    3.936890e+01 -8.744069e+01  0.000000e+00  4.400000e+01
75%    4.195626e+01 -8.024511e+01  0.000000e+00  5.700000e+01
max    6.751027e+01 -6.695090e+01  1.000000e+00  9.600000e+01
```

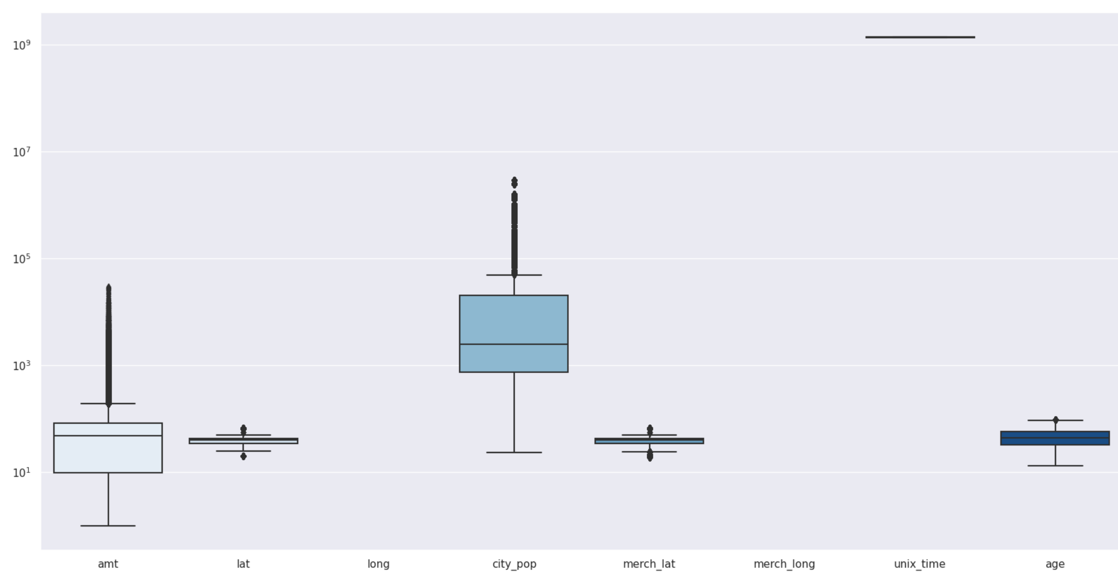
[ ]:
```
fraud.isna().values.sum()
```

[ ]: 0

There are no missing values in the dataset.

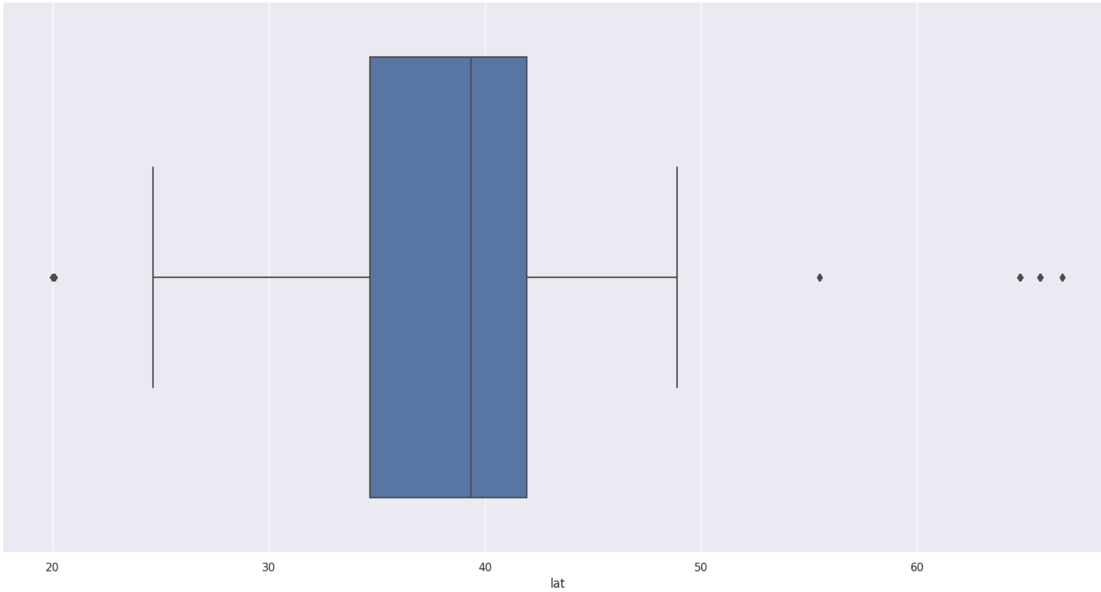
[ ]:
```
# Detecting Outliers

sns.set(rc={'figure.figsize': (20,10)})
sns.boxplot(data=fraud[['amt','lat', 'long','city_pop','merch_lat',␣
 ↪'merch_long','unix_time', 'age']], orient='v',palette='Blues')
plt.semilogy()
```

[ ]: []



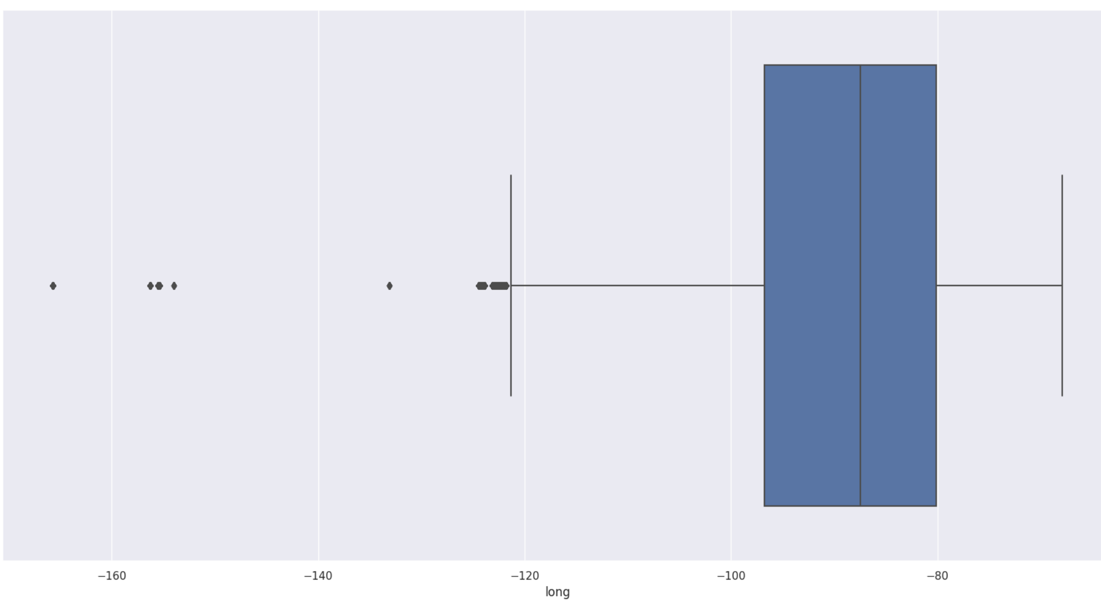

[ ]:
```
sns.boxplot(x=fraud['lat'])
```

[ ]: <Axes: xlabel='lat'>



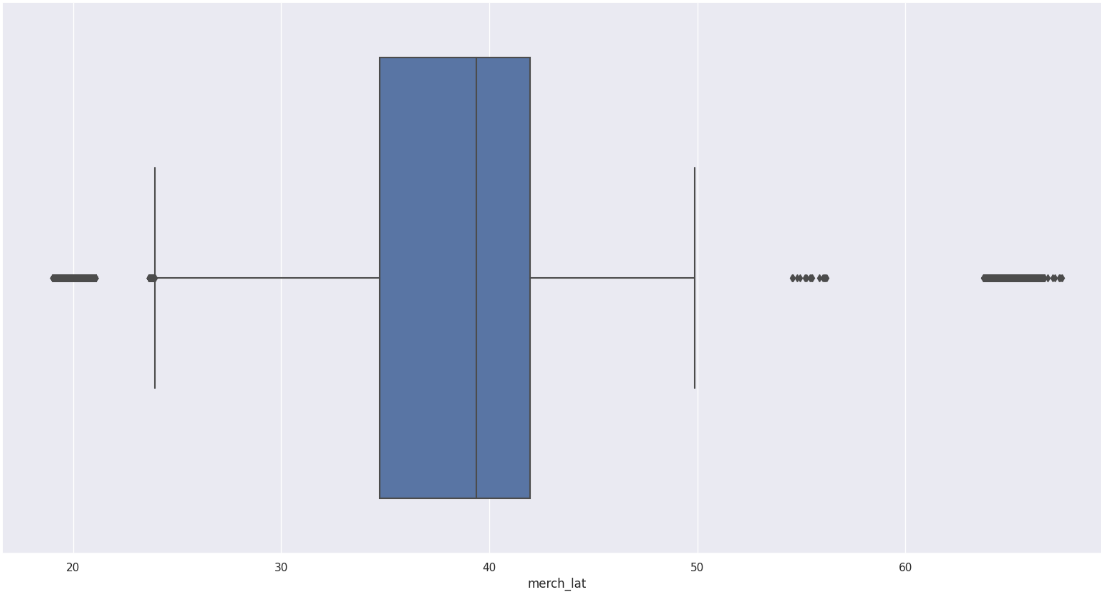

[ ]:
```
sns.boxplot(x=fraud['long'])
```

[ ]: <Axes: xlabel='long'>



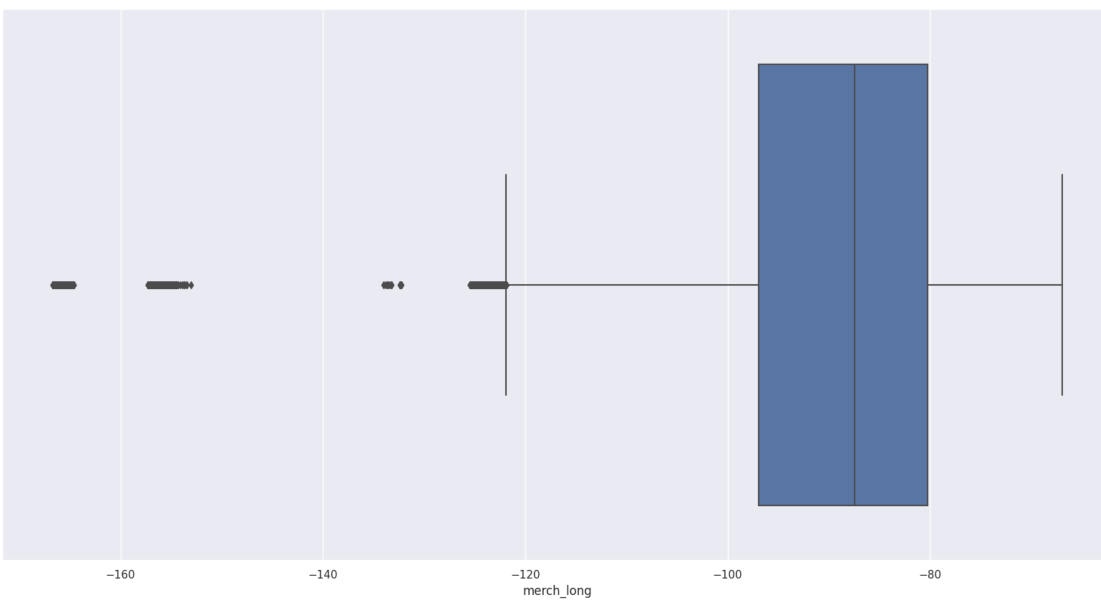

[ ]:
```
sns.boxplot(x=fraud['merch_lat'])
```

[ ]: <Axes: xlabel='merch_lat'>



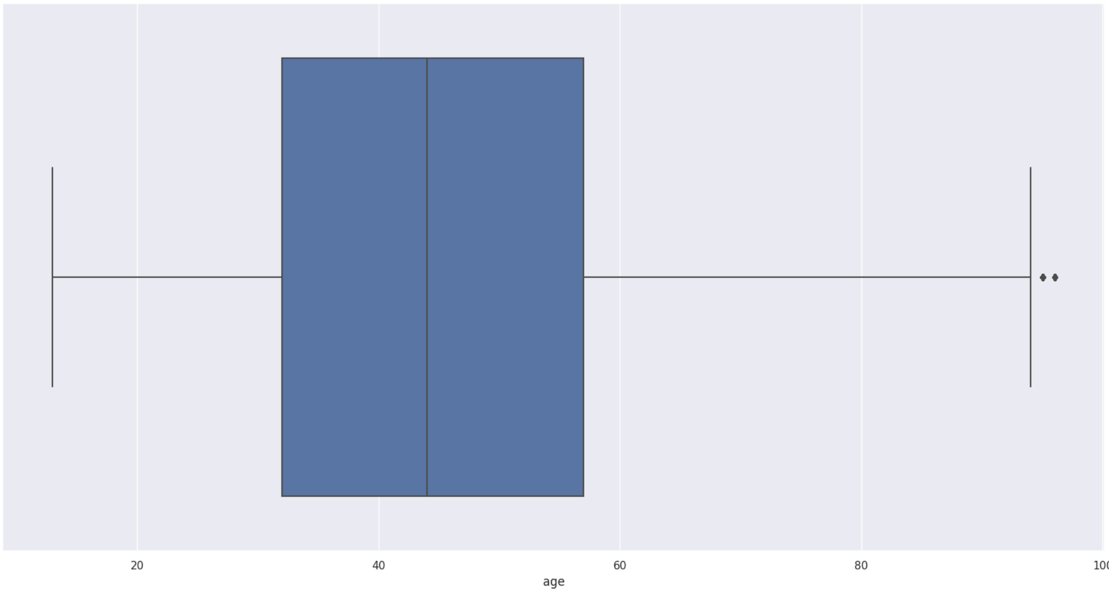

```
[ ]: sns.boxplot(x=fraud['merch_long'])
```

[ ]: <Axes: xlabel='merch_long'>




```
[ ]: sns.boxplot(x=fraud['age'])
```

[ ]: <Axes: xlabel='age'>




```
[ ]: def locate_outliers(fraud):
       Q1 = fraud.quantile(0.25)
       Q3 = fraud.quantile(0.75)
       IQR = Q3 - Q1
       outliers = fraud[((fraud < (Q1 - 1.5 * IQR)) | (fraud > (Q3 + 1.5 * IQR)))]
       return outliers
```

```
[ ]: outliers = locate_outliers(fraud[['amt', 'lat', 'long', 'city_pop',␣
      ↪'merch_lat', 'merch_long', 'age']])
     outliers.count()
```

```
[ ]: amt            95054
     lat             6612
     long           71026
     city_pop      346191
     merch_lat       7063
     merch_long     59972
     age              455
     dtype: int64
```

Observing the boxplots, it is clear that each of the numeric features contain outliers.

```
[ ]: fraud.nunique()
```

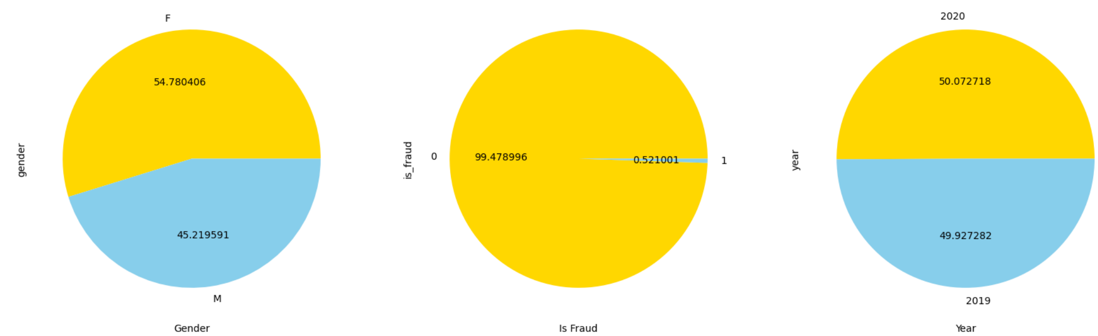
```
[ ]: cc_num                1
     merchant            693
     category             14
     amt               60616
```

```
first             355
last              486
gender              2
street            999
city              906
state              51
zip                 1
lat               983
long              983
city_pop          891
job               497
trans_num     1852394
unix_time     1819583
merch_lat     1754157
merch_long    1809753
is_fraud            2
month              12
day                 7
year                2
age                84
dtype: int64
```

The values above displays the number of unique values for each variable.

Checking the distribution of variables with binary values

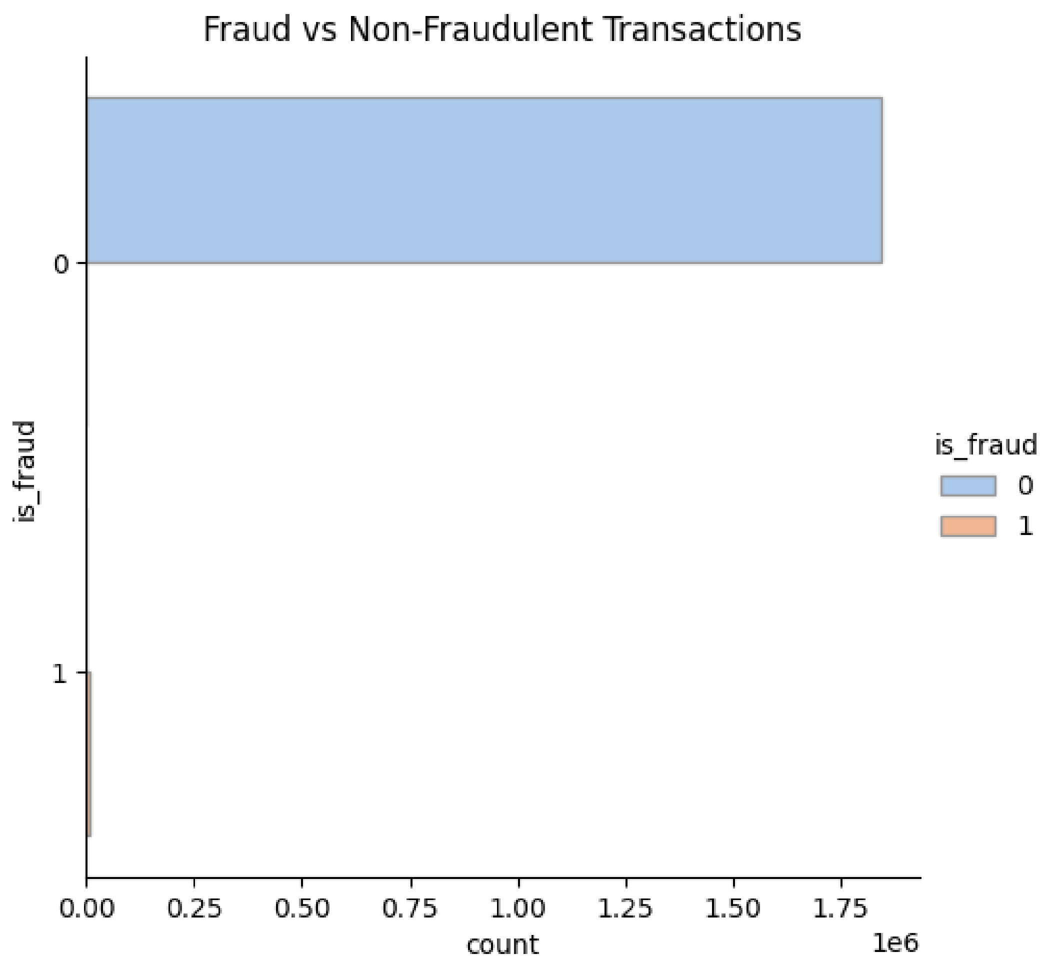
[ ]:
```
fig, (ax1, ax2, ax3) = plt.subplots(ncols=3, figsize=(20,30))
fraud['gender'].value_counts().plot(kind='pie', autopct='%2f', colors=['gold',␣
↪'skyblue'], ax=ax1)
fraud['is_fraud'].value_counts().plot(kind='pie', autopct='%2f',␣
↪colors=['gold', 'skyblue'], ax=ax2)
fraud['year'].value_counts().plot(kind='pie', autopct='%2f', colors=['gold',␣
↪'skyblue'], ax=ax3)

ax1.set_xlabel('Gender')
ax2.set_xlabel('Is Fraud')
ax3.set_xlabel('Year')
plt.show()
```




The above pie charts illustrate the proportion of the binary values for gender, is_fraud, and year. For gender, 54.78% of observations in the data set are females while the remaining 45.22% are males. As for fraud, there is a clear imbalance between the 2 classes where 99.48% of the transactions are classified as not fraudulent whereas only 0.52% are fraudulent. Therefore, a resampling technique(s) needs to be applied to overcome the issue of imbalance. Lastly, as mentioned earlier, the distribution for year is almost equal, indicating equal representation of the years in the data set.

[ ]:
```
sns.catplot(
    data = fraud, y='is_fraud', hue='is_fraud', kind='count',
    palette="pastel", edgecolor=".6"
).set(title = 'Fraud vs Non-Fraudulent Transactions')
```

[ ]: <seaborn.axisgrid.FacetGrid at 0x7f3033b4e6e0>




Again, but in the form of a horizontal barplot, this shows the unequal distribution between legitmate and fraudulent transactions and we can still see that majority of the transactions are classified as legitimate.

[ ]:
```
top_10_city = fraud['city'].value_counts().sort_values(ascending=False)
top_10_city = top_10_city.head(10)
top_10_city
```

[ ]:
```
Birmingham     8040
San Antonio    7312
Utica          7309
Phoenix        7297
Meridian       7289
Warren         6584
Conway         6574
Cleveland      6572
Thomas         6571
```

```
Houston        5865
Name: city, dtype: int64
```

[ ]:
```
top_10_states = fraud['state'].value_counts().sort_values(ascending=False)
top_10_states = top_10_states.head(10)
top_10_states
```

[ ]:
```
TX    135269
NY    119419
PA    114173
CA     80495
OH     66627
MI     65825
IL     62212
FL     60775
AL     58521
MO     54904
Name: state, dtype: int64
```

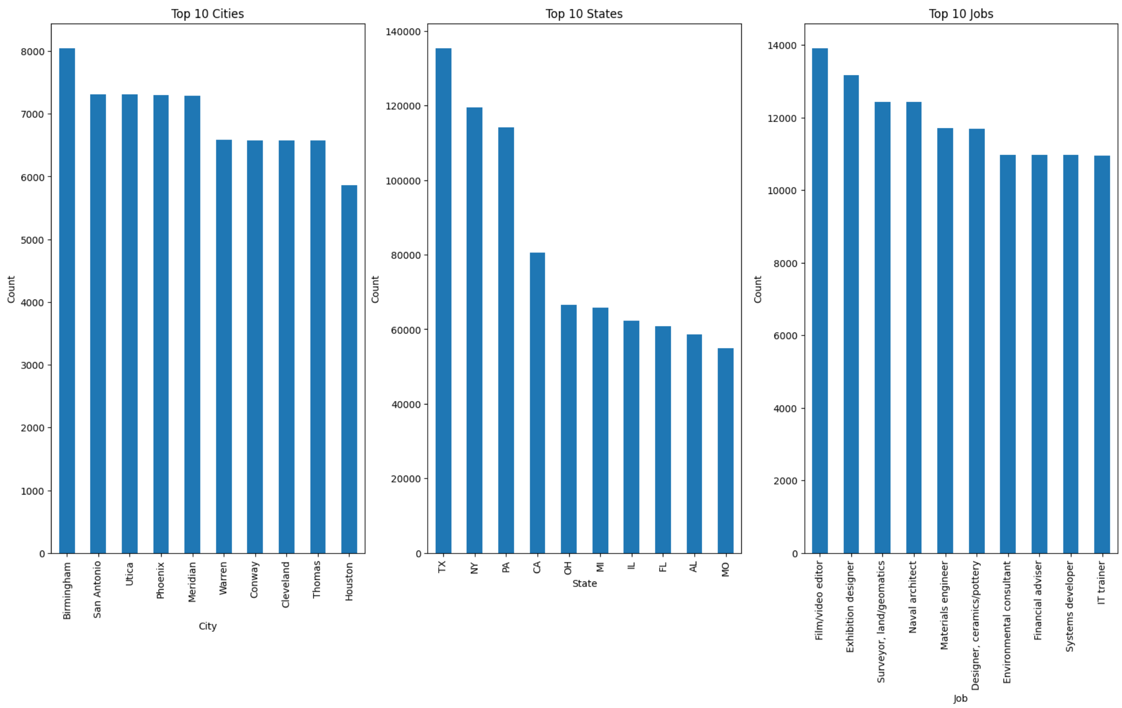
[ ]:
```
top_10_jobs = fraud['job'].value_counts().sort_values(ascending=False)
top_10_jobs = top_10_jobs.head(10)
top_10_jobs
```

[ ]:
```
Film/video editor              13898
Exhibition designer            13167
Surveyor, land/geomatics       12436
Naval architect                12434
Materials engineer             11711
Designer, ceramics/pottery     11688
Environmental consultant       10974
Financial adviser              10963
Systems developer              10962
IT trainer                     10943
Name: job, dtype: int64
```

Given that city, state, and job have large unique values (city: 906, state: 51, job: 497), it would be difficult to visualize each in a barplot. Therefore, to overcome this, I took the top 10 cities, states, and jobs based on their frequency to ease the analysis.

[ ]:
```
fig, ax = plt.subplots(1,3,figsize=(20,10))
top_10_city.plot(kind='bar', ax=ax[0]).set_title('Top 10 Cities')
top_10_states.plot(kind='bar', ax=ax[1]).set_title('Top 10 States')
top_10_jobs.plot(kind='bar', ax=ax[2]).set_title('Top 10 Jobs')

ax[0].set_xlabel('City')
ax[1].set_xlabel('State')
ax[2].set_xlabel('Job')
```

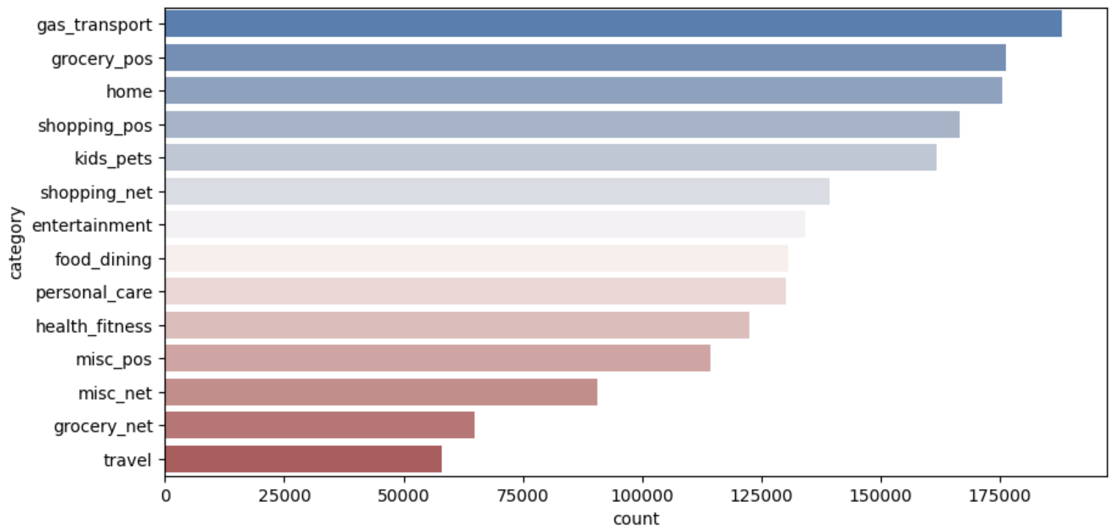
```
ax[0].set_ylabel('Count')
ax[1].set_ylabel('Count')
ax[2].set_ylabel('Count')
plt.show()
```

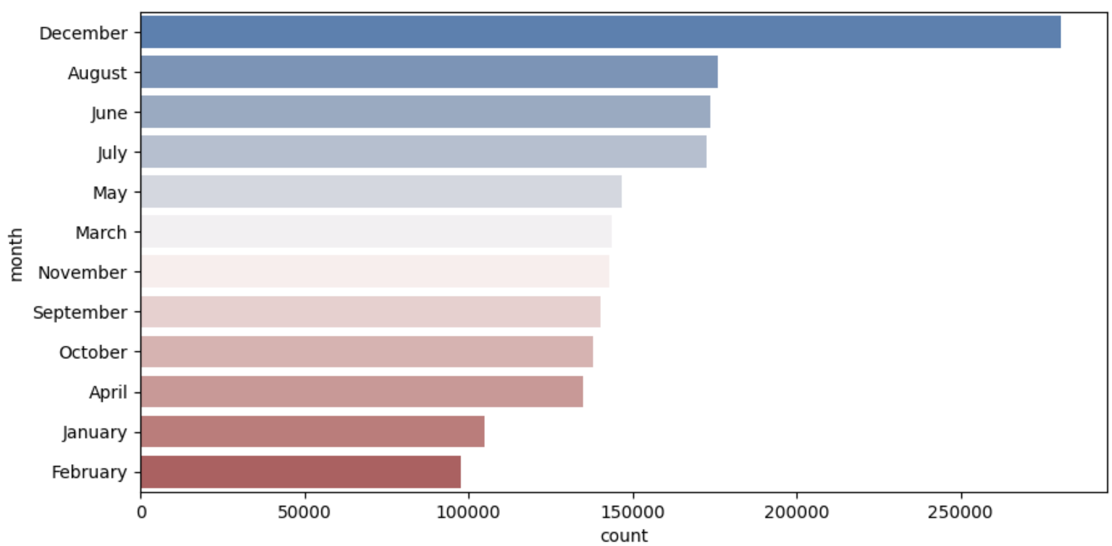


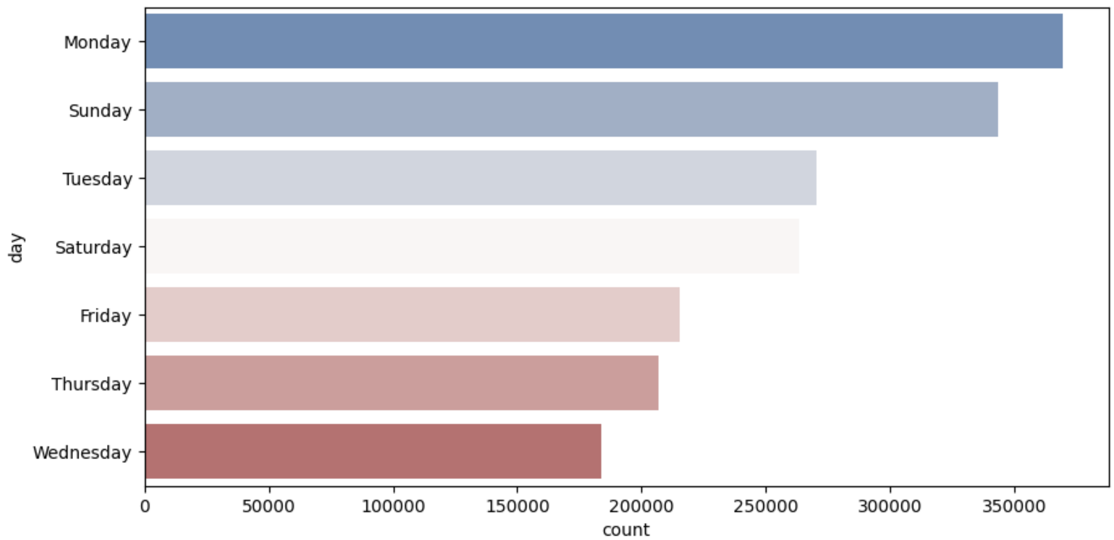

The above bar plots displays each of the top 10 observations for city, state, and job. Looking at each plot, starting with city, majority of the transactions took place in Birmingham whereas for state, Texas (TX) stood out to be the majority class in the variable and is the state where most transactions have occured. For job, a large number of credit card holder have jobs as film/video editors.

[ ]:
```
cat_features = ['category', 'month', 'day'] # Selecting a few of the␣
↪categorical features
num_features = ['amt', 'lat', 'long', 'city_pop', 'unix_time', 'merch_lat',␣
↪'merch_long', 'age'] # Numeric features
```

[ ]:
```
for i in cat_features:
  fig, ax = plt.subplots(1,1, figsize=(10,5))
  sns.countplot(y=fraud[i][1:], data=fraud.iloc[1:], order=fraud[i][1:].
↪value_counts().index, palette='vlag')
  plt.yticks(fontsize=10)
  plt.xticks(fontsize=10)
plt.show()
```

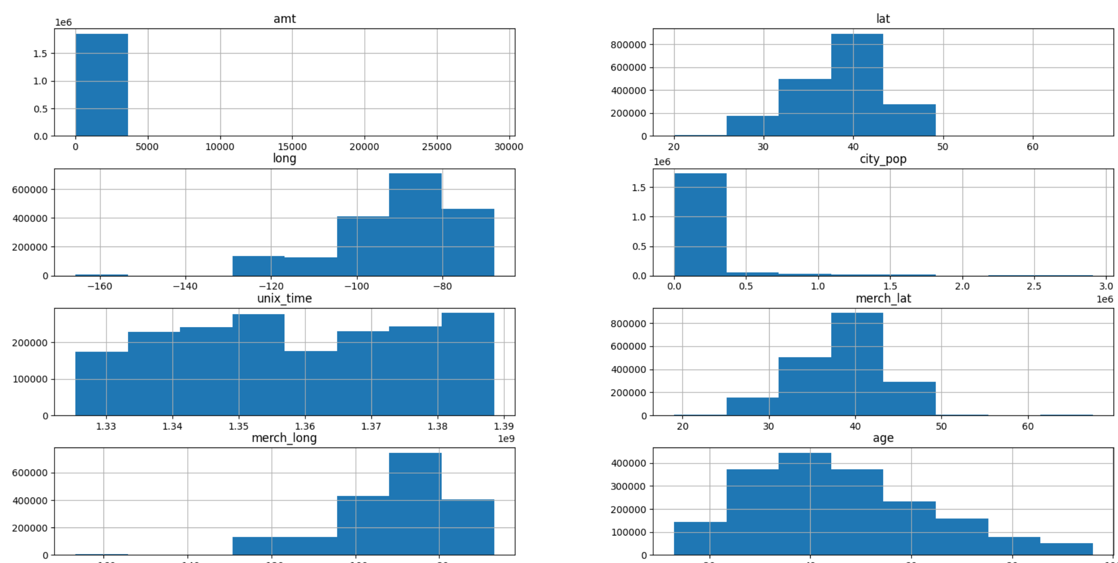

gas_transport
grocery_pos
home
shopping_pos
kids_pets
shopping_net
entertainment
food_dining
personal_care
health_fitness
misc_pos
misc_net
grocery_net
travel
category
0
25000
50000
75000
100000
125000
150000
175000
count


December
August
June
July
May
March
November
September
October
April
January
February
month
0
50000
100000
150000
200000
250000
count




Analyzing a few of the other categorical variables, we see that for the category of transaction, most of the individuals used their credit cards for gas and transportation, followed by groceries and home. As discussed earlier, most transactions were performed during the month of December and on a Monday.

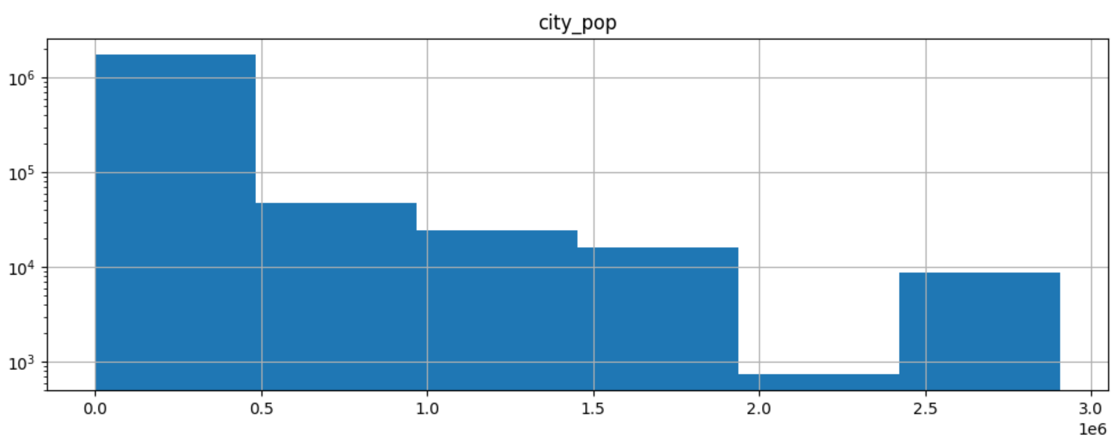
[ ]:
```
fig, ax = plt.subplots(4,2, figsize=(20,10))
for ax, c in zip(ax.flatten(), num_features):
  fraud.hist(column=c, ax=ax, bins=8)
plt.show()
```



Examining the distributions of the numeric attributes through a histogram, what can generalized is that most of the features are skewed. For both longitude and latitude of the transactions, they are each skewed in opposite in directions, latitude being right-skewed and longitude being left skewed and exactly the same interpretation can be made for merchant latitude and longitude. Unix_time on the other hand is the only numerical feature where there exists no skew. For age, it can be inferred that a large number of the individuals in the data are between the ages of 20 and 40 years old and looking at its distribution, it is right skewed. The distribution for amt (amount) and city_pop is not clear in this group plot therefore, plotting each of these in terms of logs will be required.

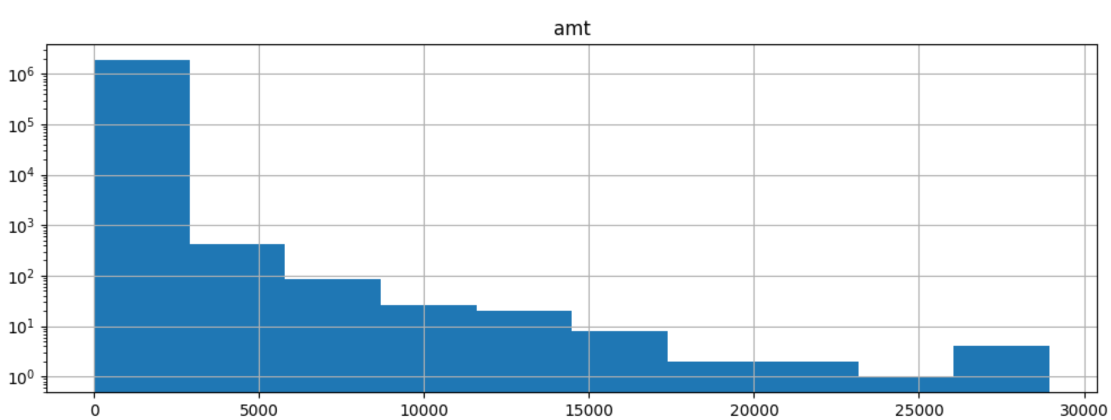
[ ]:
```
fraud.hist(column=['city_pop'], figsize=(12,4), bins=6)
plt.semilogy()
```

[ ]: []




[ ]:
```
fraud.hist(column=['amt'], figsize=(12,4))
plt.semilogy()
```

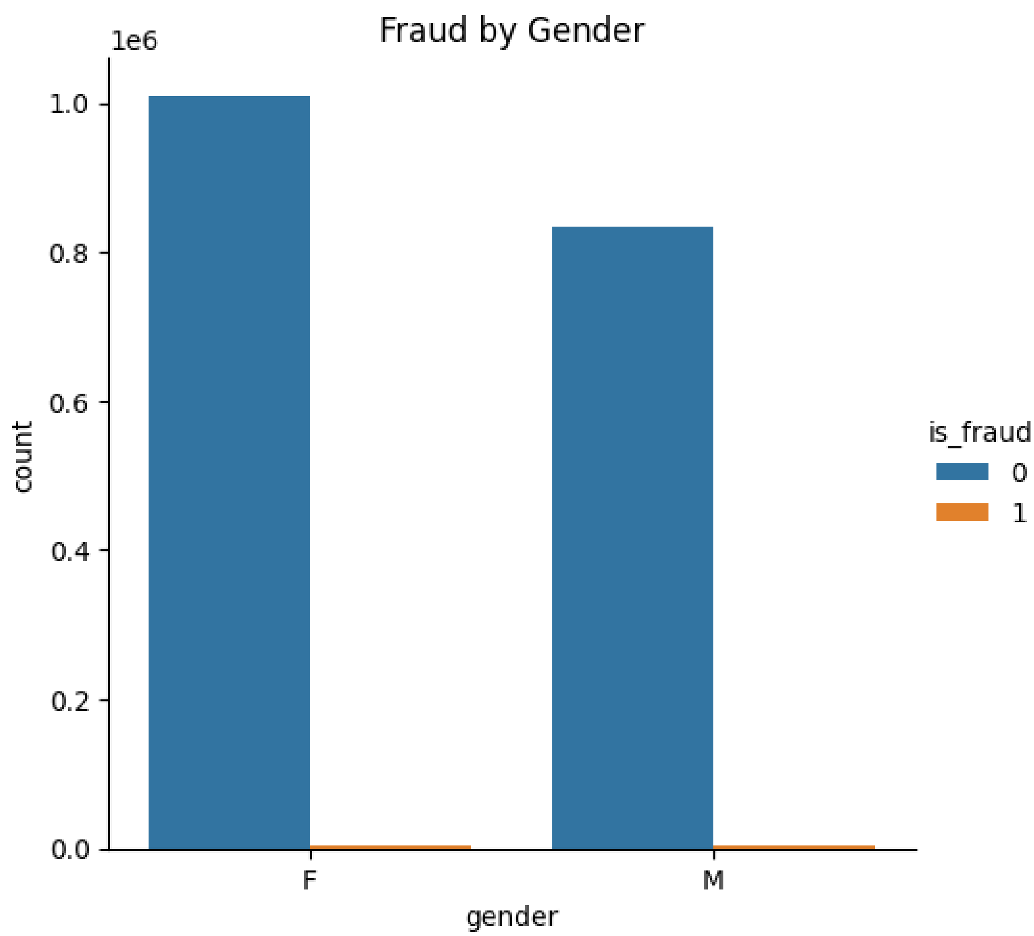
[ ]: []



Getting a better visualiztion of the distribution of city_pop and amt, we can see that each of their distributions are right skewed.

[ ]:
```
sns.catplot(
    data = fraud, x='gender', hue='is_fraud', kind='count'
).set(title='Fraud by Gender')
```

[ ]: <seaborn.axisgrid.FacetGrid at 0x7f2f76394eb0>




In the barplot above, both males and females in the data for the most part have not made fraudulent transactions. However, there appeares to be a few observations that are fraudulent for both genders though its not easlity visible on this graph.

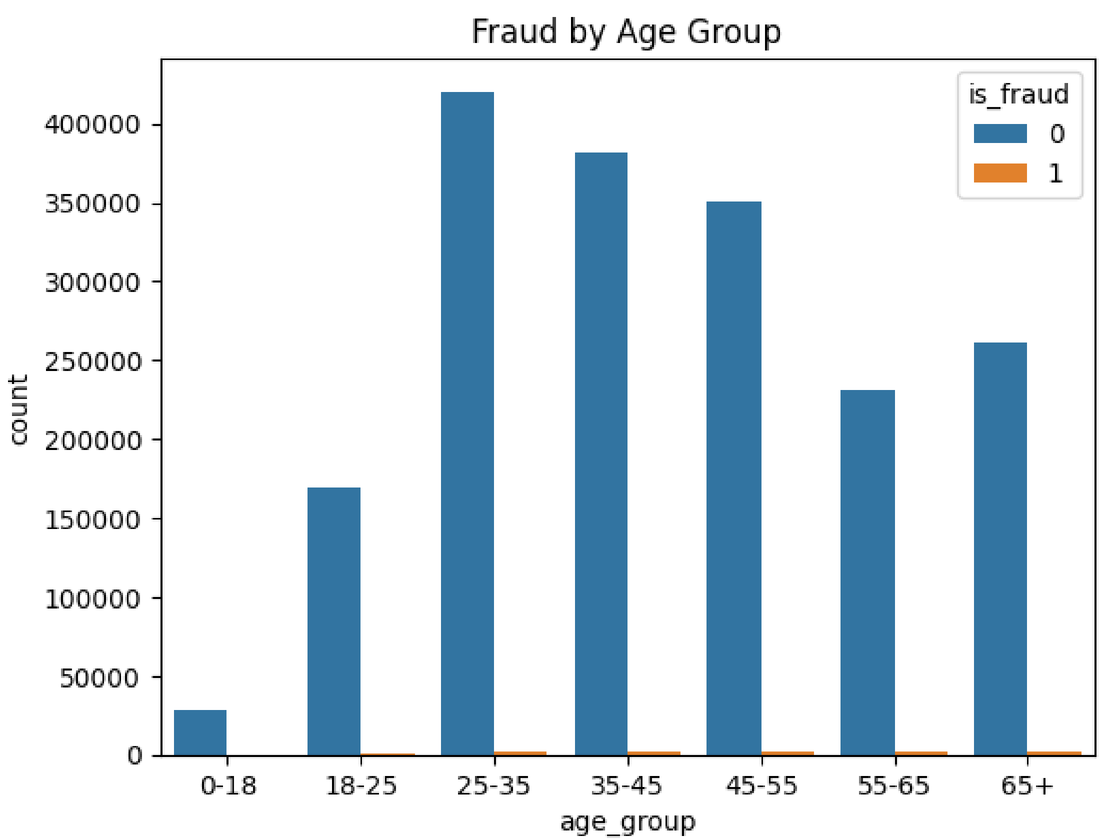
[ ]:
```
fraud.groupby(fraud['gender'])['is_fraud'].value_counts()
```

```
[ ]: gender  is_fraud
     F       0           1009850
             1              4899
     M       0            832893
             1              4752
     Name: is_fraud, dtype: int64
```

Taking a closer look, though the frequency is pretty small in comparison to non-fraud, females are making more fraud transactions than males however, the difference between the two genders is not large.

```
[ ]: fraud['age_group'] = pd.cut(fraud.age, bins=[0,18,25,35,45,55,65,float('inf')],␣
      ↪right=True, labels=['0-18', '18-25', '25-35', '35-45', '45-55', '55-65',␣
      ↪'65+'])
     sns.countplot(data=fraud, x='age_group', hue='is_fraud').set_title('Fraud by␣
      ↪Age Group')
```

```
[ ]: Text(0.5, 1.0, 'Fraud by Age Group')
```



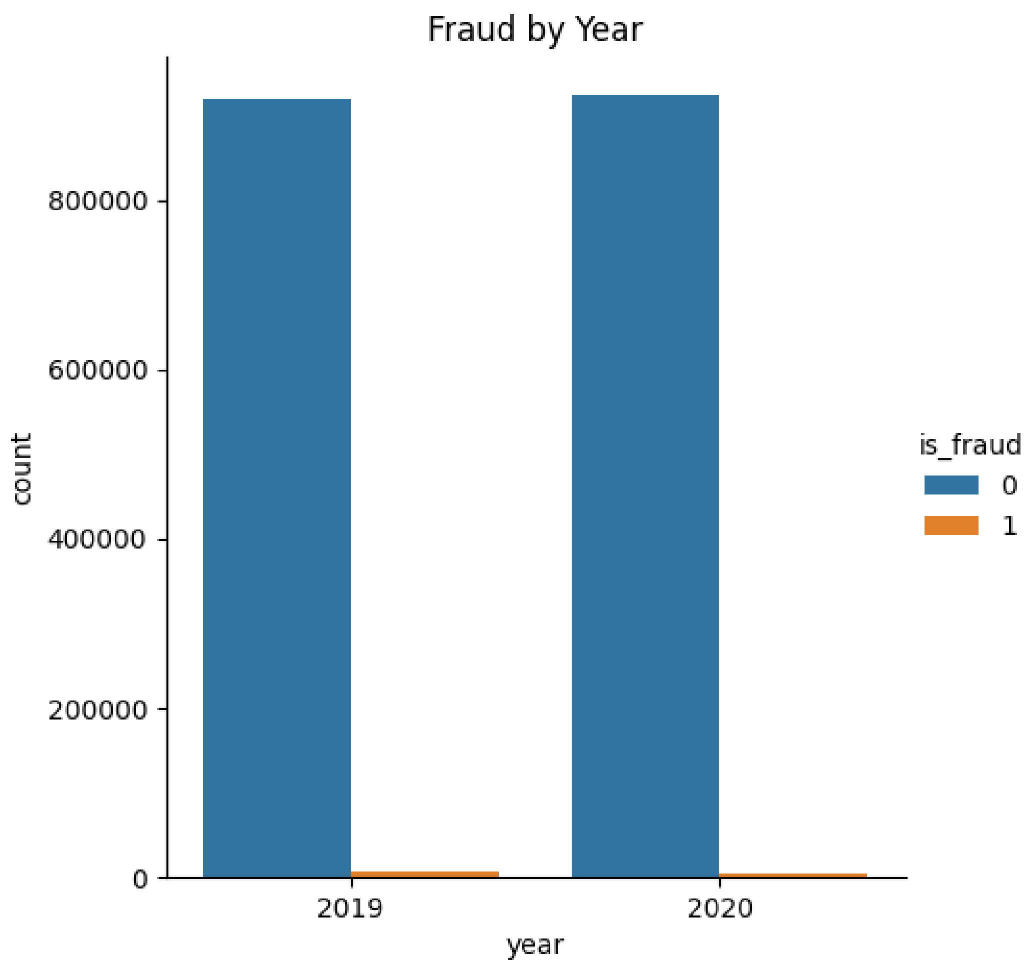

```
[ ]: fraud.groupby(fraud['age_group'])['is_fraud'].value_counts()
```

[ ]:
```
age_group  is_fraud
0-18       0            27753
           1              149
18-25      0           169647
           1              951
25-35      0           420229
           1             1830
35-45      0           381961
           1             1522
45-55      0           350912
           1             1850
55-65      0           231492
           1             1585
65+        0           260749
           1             1764
Name: is_fraud, dtype: int64
```

With respect to age range of credit card holders involved in fraudulent transactions, people between the ages 45-55 had the highest occurence of fraud, followed by individuals between the ages 25-35, 65+, 55-65, and 35-45 years of age.

[ ]:
```
sns.catplot(
    data = fraud, x='year',hue='is_fraud', kind='count'
).set(title='Fraud by Year')
```

[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x7f2fc5654790>
```

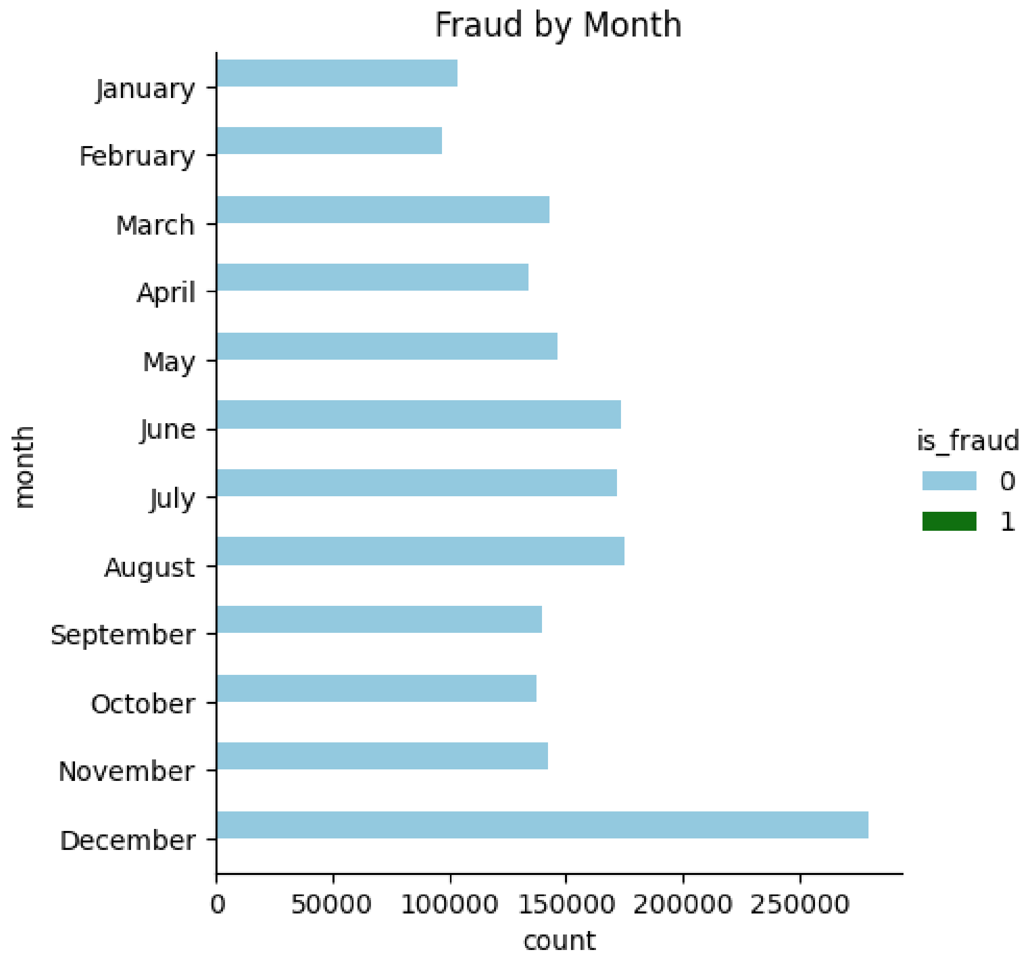



[ ]:
```
fraud.groupby(fraud['year'])['is_fraud'].value_counts()
```

[ ]:
```
year  is_fraud
2019  0           919630
      1             5220
2020  0           923113
      1             4431
Name: is_fraud, dtype: int64
```

Similar to fraud by gender, most individuals did not perform fraudulent transactions for both years. However, taking a further look, more fraudulent transactions were performed in 2019 than in 2020.

[ ]:
```
sns.catplot(
    data = fraud,y='month',hue='is_fraud', kind='count', palette=sns.
 ↪color_palette(['skyblue', 'green'])
).set(title='Fraud by Month')
plt.rcParams['figure.figsize']=(10,20)
```




[ ]:
```
fraud.groupby(fraud['month'])['is_fraud'].value_counts()
```

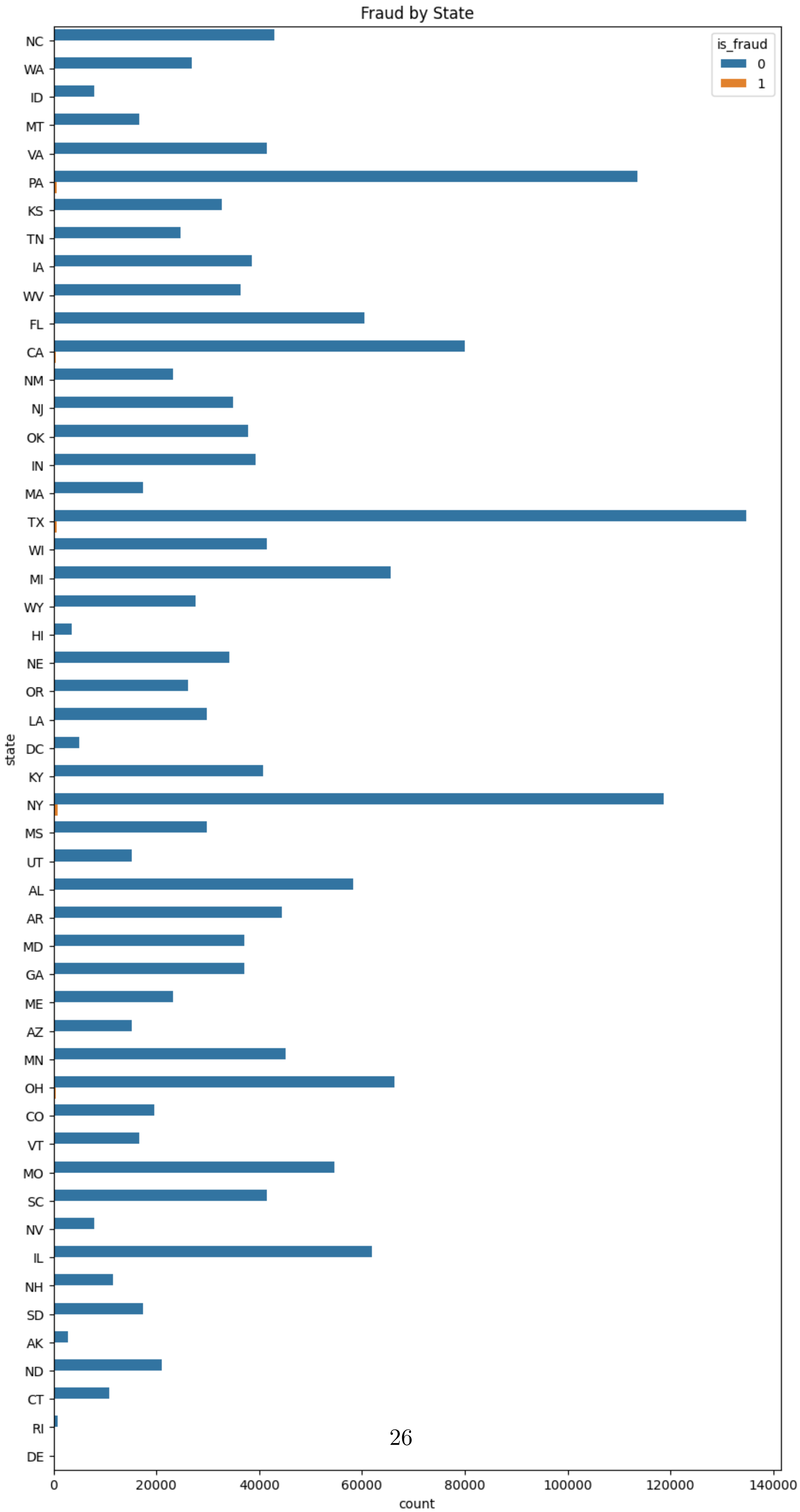
[ ]:
```
month      is_fraud
April      0           134292
           1              678
August     0           175321
           1              797
December   0           279748
           1              850
February   0            96804
           1              853
January    0           103878
           1              849
July       0           171792
           1              652
June       0           173048
           1              821
```

```
March      0           142851
           1              938
May        0           145940
           1              935
November   0           142374
           1              682
October    0           137268
           1              838
September  0           139427
           1              758
Name: is_fraud, dtype: int64
```

Though the barplot above shows that no fraudulent transaction were made for each month, when breaking it down, the above output shows that the month of March had the highest number of fraudulent transactions in comparison to all the months, followed by May, Feburary, December, and January.

[ ]:
```
sns.countplot(data=fraud, y='state', hue='is_fraud').set(title='Fraud by State')
```

[ ]:
```
[Text(0.5, 1.0, 'Fraud by State')]
```



Fraud by State
is_fraud
0
1
state
count
NC
WA
ID
MT
VA
PA
KS
TN
IA
WV
FL
CA
NM
NJ
OK
IN
MA
TX
WI
MI
WY
HI
NE
OR
LA
DC
KY
NY
MS
UT
AL
AR
MD
GA
ME
AZ
MN
OH
CO
VT
MO
SC
NV
IL
NH
SD
AK
ND
CT
RI
DE
0
20000
40000
60000
80000
100000
120000
140000

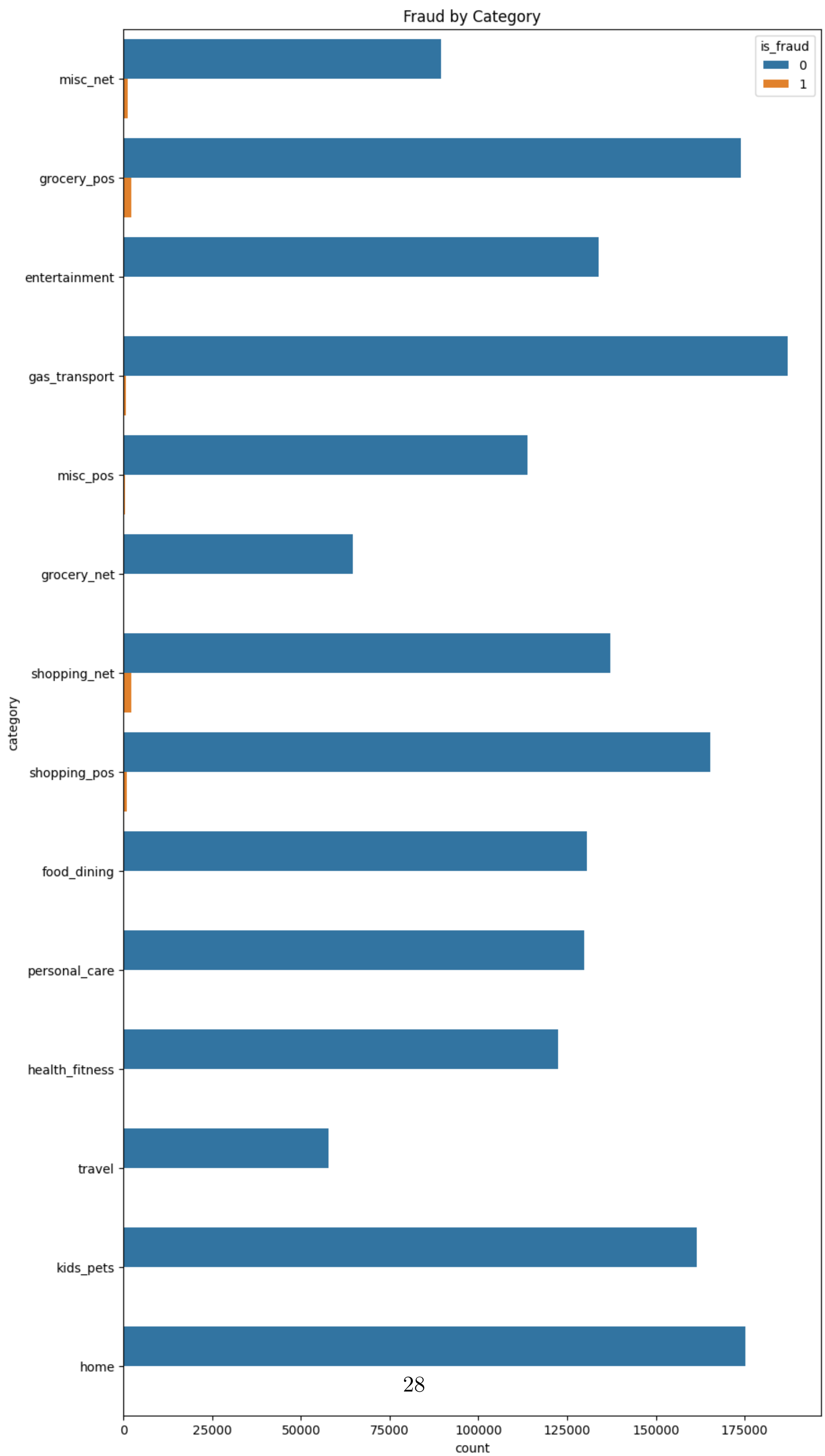
[ ]:
```
fraud.groupby(fraud['is_fraud'])['state'].value_counts()
```

[ ]:
```
is_fraud  state
0         TX       134677
          NY       118689
          PA       113601
          CA        80093
          OH        66267
                     …
1         ID           33
          DC           31
          HI           16
          RI           15
          DE            9
Name: state, Length: 101, dtype: int64
```

In the barplot, it can be observed that there are a few states that visbly show that fraudulent transaction have occured in those states, namely New York State (NY), Pennsylvania (PA), Virginia (VA), Texas (TX) and others. Again, like with the previous analysis, most of the transactions in each of the states were not fraudulent.

[ ]:
```
sns.countplot(data=fraud, y='category', hue='is_fraud').set(title='Fraud by␣
↪Category')
```

[ ]:
```
[Text(0.5, 1.0, 'Fraud by Category')]
```


Fraud by Category
is_fraud
0
1
misc_net
grocery_pos
entertainment
gas_transport
misc_pos
grocery_net
shopping_net
shopping_pos
food_dining
personal_care
health_fitness
travel
kids_pets
home
category
0
25000
50000
75000
100000
125000
150000
175000
count

The countplot above shows that out of all the categories people made transactions on, shopping, miscellaneous, and gas and transportation appeared to have higher counts of fraud as opposed to the remaining categories.

[ ]:
```
fraud = fraud.drop('age_group', axis=1)
```

Since there are 2 age variables, I removed age_group since it will no longer be needed for the remainder of the analysis

[ ]:
```
#pip install pandas-profiling
```

[ ]:
```
from pandas_profiling import ProfileReport
profile = ProfileReport(fraud)
```

```
<ipython-input-51-65f5ce699e0f>:1: DeprecationWarning: `import pandas_profiling`
is going to be deprecated by April 1st. Please use `import ydata_profiling`
instead.
  from pandas_profiling import ProfileReport
```

[ ]:
```
profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

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

The Pandas profile report above summarizes the dataset and some of the analysis I have conducted.

***Data Cleaning (returned)***

[12]:
```
fraud = fraud.drop(['cc_num', 'first', 'last', 'zip', 'trans_num',␣
 ↪'unix_time'], axis = 1)
```

After getting a better understanding of the dataset, it was realized that some of these variables do not contain valuable information in regards to understanding what is considered a fraudulent transaction. Namely, these are cc_num, first, last, zip, trans_num, and unix_time. For cc_num and zip, they both have 1 unique value, meaning that each of the transactions have the same value, hence not contributing useful information when creating the models. As well, first and last name of the credit card holder is considered confidential and therefore would need to be reomoved from the dataset.

Transforming categorical features

[13]:
```
keep_city = fraud['city'].value_counts().index[:10]
fraud['city'] = np.where(fraud['city'].isin(keep_city), fraud['city'], 'Other' )
```

```
fraud['city'].value_counts()
```

[13]:
```
Other           1782981
Birmingham         8040
San Antonio        7312
Utica              7309
Phoenix            7297
Meridian           7289
Warren             6584
Conway             6574
Cleveland          6572
Thomas             6571
Houston            5865
Name: city, dtype: int64
```

[14]:
```
keep_state = fraud['state'].value_counts().index[:10]
fraud['state'] = np.where(fraud['state'].isin(keep_state), fraud['state'],␣
↪'Other')

# fraud['state'].value_counts()
```

[15]:
```
keep_job = fraud['job'].value_counts().index[:10]
fraud['job'] = np.where(fraud['job'].isin(keep_job), fraud['job'], 'Other')

#fraud['job'].value_counts()
```

[16]:
```
keep_merchant = fraud['merchant'].value_counts().index[:10]
fraud['merchant'] = np.where(fraud['merchant'].isin(keep_merchant),␣
↪fraud['merchant'], 'Other')

#fraud['merchant'].value_counts()
```

[17]:
```
keep_street = fraud['street'].value_counts().index[:10]
fraud['street'] = np.where(fraud['street'].isin(keep_street), fraud['street'],␣
↪'Other')

#fraud['street'].value_counts()
```

As mentioned earlier, each of these categorical features contain a high number of unique values in which when we one - hot enode them, the dataset will become increasingly large. Hence, to prevent this, these variables will be transformed to keep the top 10 levels based on their frequency and will assign the remaining levels to a level called "Other".

[18]:
```
# Converting gender and year variable to their binary values
for col in ['gender', 'year']:
    fraud[col] = fraud[col].astype('category')
    fraud[col] = fraud[col].cat.codes
```

Here, 0 for the gender variable is female and 1 being male whereas for year, 0 represents 2019 and 1 being 2020.

Dealing with outliers

[19]:
```
# Copying the original dataset
fraud_copy = fraud.copy()
```

[20]:
```
# Transforming outliers using iqr method
trans = RobustScaler()
fraud_trans = fraud
fraud_trans[['amt', 'lat', 'long', 'city_pop', 'merch_lat', 'merch_long',␣
↪'age']] = trans.fit_transform(fraud[['amt', 'lat', 'long', 'city_pop',␣
↪'merch_lat', 'merch_long', 'age']])
```

[21]:
```
fraud_trans.describe()
```

[21]:
```
                amt        gender           lat          long      city_pop  \
count  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06
mean   3.078351e-01  4.521959e-01 -1.120799e-01 -1.653205e-01  4.400913e+00
std    2.167901e+00  4.977097e-01  6.974449e-01  8.261956e-01  1.539223e+01
min   -6.323169e-01  0.000000e+00 -2.657939e+00 -4.699243e+00 -1.235513e-01
25%   -5.147019e-01  0.000000e+00 -6.443512e-01 -5.601623e-01 -8.689437e-02
50%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%    4.852981e-01  1.000000e+00  3.556488e-01  4.398377e-01  9.131056e-01
max    3.934311e+02  1.000000e+00  3.759747e+00  1.173474e+00  1.482747e+02

          merch_lat    merch_long      is_fraud          year           age
count  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06  1.852394e+06
mean  -1.150095e-01 -1.673586e-01  5.210015e-03  5.007272e-01  7.187585e-02
std    7.075255e-01  8.261930e-01  7.199217e-02  4.999996e-01  6.969570e-01
min   -2.818886e+00 -4.757374e+00  0.000000e+00  0.000000e+00 -1.240000e+00
25%   -6.414479e-01 -5.679451e-01  0.000000e+00  0.000000e+00 -4.800000e-01
50%    4.923288e-16  4.266414e-16  0.000000e+00  1.000000e+00  0.000000e+00
75%    3.585521e-01  4.320549e-01  0.000000e+00  1.000000e+00  5.200000e-01
max    3.899781e+00  1.230298e+00  1.000000e+00  1.000000e+00  2.080000e+00
```

Data without outliers

[22]:
```
def outliers(df, feature):
    Q1 = df[feature].quantile(0.25)
    Q3 = df[feature].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
```

```
    ls = df.index[(df[feature] < lower_bound) | (df[feature] > upper_bound)] #␣
↪stores indexes of outliers

    return ls
```

[23]:
```
index_list = []
for feat in ['amt', 'lat', 'long', 'city_pop', 'merch_lat', 'merch_long',␣
↪'age']:
    index_list.extend(outliers(fraud_copy, feat))
```

[24]:
```
def remove(df, ls):
    ls = sorted(set(ls))
    df = df.drop(ls)
    return df
```

[25]:
```
fraud_no_outliers = remove(fraud_copy, index_list)
fraud_no_outliers.shape[0]
```

[25]: 1174016

[26]:
```
# Scaling the data using MinMax
scaler = MinMaxScaler()
fraud_no_outliers[['amt', 'lat', 'long', 'city_pop', 'merch_lat', 'merch_long',␣
↪'age']] = scaler.fit_transform(fraud_no_outliers[['amt', 'lat', 'long',␣
↪'city_pop', 'merch_lat', 'merch_long', 'age']])
```

[27]:
```
fraud_no_outliers.describe() # Check
```

[27]:

| | amt | gender | lat | long | city_pop |
|---|---|---|---|---|---|
| count | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 |
| mean | 2.602895e-01 | 4.579520e-01 | 5.928507e-01 | 6.141770e-01 | 1.064837e-01 |
| std | 2.268896e-01 | 4.982290e-01 | 1.896967e-01 | 2.156431e-01 | 1.850823e-01 |
| min | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 4.326798e-02 | 0.000000e+00 | 4.601954e-01 | 4.788225e-01 | 1.158373e-02 |
| 50% | 2.293931e-01 | 0.000000e+00 | 6.241968e-01 | 6.444941e-01 | 3.400386e-02 |
| 75% | 3.967445e-01 | 1.000000e+00 | 7.234000e-01 | 7.778613e-01 | 1.011397e-01 |
| max | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 |

| | merch_lat | merch_long | is_fraud | year | age |
|---|---|---|---|---|---|
| count | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 | 1.174016e+06 |
| mean | 5.816226e-01 | 6.090959e-01 | 1.356029e-03 | 5.010366e-01 | 4.123622e-01 |
| std | 1.783949e-01 | 2.085860e-01 | 3.679934e-02 | 4.999991e-01 | 2.176340e-01 |
| min | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 |
| 25% | 4.502774e-01 | 4.759202e-01 | 0.000000e+00 | 0.000000e+00 | 2.345679e-01 |
| 50% | 6.097821e-01 | 6.348946e-01 | 0.000000e+00 | 1.000000e+00 | 3.827160e-01 |
| 75% | 7.025831e-01 | 7.701973e-01 | 0.000000e+00 | 1.000000e+00 | 5.555556e-01 |
| max | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 |

One-hot-encoding categorical features (**dataset with transformed outliers**)

[28]:
```
cat_columns = ['merchant', 'category', 'street', 'city', 'state', 'job',
    'month', 'day']
```

[29]:
```
fraud_trans = pd.get_dummies(fraud_trans, columns = cat_columns, prefix =
    cat_columns)
fraud_trans.head(5)
```

[29]:
```
        amt  gender       lat      long  city_pop  merch_lat  merch_long  \
0 -0.578274       0 -0.450457  0.378534  0.053709  -0.465291    0.323782
1  0.813776       0  1.311077 -1.846971 -0.117118   1.356701   -1.846112
2  2.350395       1  0.388709 -1.489489  0.087354   0.524076   -1.483925
3 -0.033351       1  0.945651 -1.480583 -0.025731   1.062262   -1.508339
4 -0.074735       1 -0.128392  0.481611 -0.119671  -0.096160    0.528885

   is_fraud  year   age  …  month_November  month_October  month_September  \
0         0     0 -0.56  …               0              0                0
1         0     0 -0.16  …               0              0                0
2         0     0  0.48  …               0              0                0
3         0     0  0.32  …               0              0                0
4         0     0 -0.48  …               0              0                0

   day_Friday  day_Monday  day_Saturday  day_Sunday  day_Thursday  \
0           0           0             0           0             0
1           0           0             0           0             0
2           0           0             0           0             0
3           0           0             0           0             0
4           0           0             0           0             0

   day_Tuesday  day_Wednesday
0            1              0
1            1              0
2            1              0
3            1              0
4            1              0

[5 rows x 98 columns]
```

One-hot-encoding categorical features (**dataset without outliers**)

[30]:
```
fraud_no_outliers = pd.get_dummies(fraud_no_outliers, columns = cat_columns,
    prefix = cat_columns)
fraud_no_outliers.head(5)
```

[30]:
```
        amt  gender       lat      long  city_pop  merch_lat  merch_long  \
1  0.552447       0  1.000000  0.052785  0.002616   0.971950    0.067213
4  0.213012       1  0.568048  0.783031  0.001578   0.568266    0.787327
```

```
5  0.486921        0  0.648697  0.863286  0.044321   0.644443    0.832474
7  0.367414        1  0.585484  0.799288  0.124452   0.578781    0.789005
8  0.017006        0  0.647084  0.779303  0.030080   0.632831    0.763192

   is_fraud  year       age  …  month_November  month_October  \
1         0     0  0.333333  …               0              0
4         0     0  0.234568  …               0              0
5         0     0  0.543210  …               0              0
7         0     0  0.716049  …               0              0
8         0     0  0.790123  …               0              0

   month_September  day_Friday  day_Monday  day_Saturday  day_Sunday  \
1                0           0           0             0           0
4                0           0           0             0           0
5                0           0           0             0           0
7                0           0           0             0           0
8                0           0           0             0           0

   day_Thursday  day_Tuesday  day_Wednesday
1             0            1              0
4             0            1              0
5             0            1              0
7             0            1              0
8             0            1              0

[5 rows x 92 columns]
```

**Preparing Data for modelling** (dataset with transformed outliers)

[31]:
```
y1 = fraud_trans.is_fraud # y1 will represent the target variable for␣
↪transformed data
```

[32]:
```
col = 'is_fraud'
X1 = fraud_trans.loc[:, fraud_trans.columns != col] # X1 will represent the␣
↪predictor variables for transformed data
```

[33]:
```
skf = StratifiedKFold(n_splits = 5, shuffle = True, random_state = 300)
```

[34]:
```
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1, test_size = 0.
↪3, random_state = 300)
```

1. Under-sampling the data

[45]:
```
# Pipeline with random under sampling and logistic regression for transformed␣
↪dataset
```

```
rus_log_pipeline1 = imbpipeline(steps = [['RandomUnderSampler',␣
↪RandomUnderSampler(sampling_strategy = 'majority',random_state = 300)],
                                      ['LogisticRegression',␣
↪LogisticRegression(random_state = 300)]])
```

[46]:
```
log_param = {
    'LogisticRegression__C': [1.0, 0.1, 0.01, 0.001]
}
log_grid = GridSearchCV(rus_log_pipeline1, param_grid = log_param, cv = skf,␣
↪scoring = 'precision', return_train_score = True)
```

[ ]:
```
log_grid.fit(X1_train, y1_train)
```

[41]:
```
y1_predict = log_grid.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[41]:
```
0.08995428147782034
```

[49]:
```
# Pipeline with random under sampling and decision tree for transformed dataset

rus_tree_pipeline1 = imbpipeline(steps = [['RandomUnderSampler',␣
↪RandomUnderSampler(sampling_strategy = 'majority',random_state = 300)],
                                      ['DecisionTree', tree.
↪DecisionTreeClassifier(random_state = 300)]])
```

[51]:
```
tree_param1 = {
    'DecisionTree__criterion': ['gini', 'entropy'],
    'DecisionTree__max_depth': [5, 10, 20, 25]
}
tree_grid1 = GridSearchCV(rus_tree_pipeline1, param_grid = tree_param1, cv =␣
↪skf, scoring = 'precision', return_train_score = True)
```

[53]:
```
tree_grid1.fit(X1_train, y1_train)
```

[53]:
```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['RandomUnderSampler',
                                        RandomUnderSampler(random_state=300,
sampling_strategy='majority')],
                                       ['DecisionTree',
DecisionTreeClassifier(random_state=300)]]),
             param_grid={'DecisionTree__criterion': ['gini', 'entropy'],
                         'DecisionTree__max_depth': [5, 10, 20, 25]},
             return_train_score=True, scoring='precision')
```

[54]:
```
y1_predict = tree_grid1.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[54]: 0.11042896185685427

[69]:
```
# Pipeline with random under sampling and KNN for transformed dataset

rus_knn_pipeline1 = imbpipeline(steps = [['RandomUnderSampler',␣
 ↪RandomUnderSampler(sampling_strategy = 'majority',random_state = 300)],
                                        ['KNN', KNeighborsClassifier()]])
```

[70]:
```
knn_param1 = {
    'KNN__n_neighbors': [3, 4, 5, 6]
}
knn_grid1 = GridSearchCV(rus_knn_pipeline1, param_grid = knn_param1, cv = skf,␣
 ↪scoring = 'precision', return_train_score = True)
```

[71]:
```
knn_grid1.fit(X1_train, y1_train)
```

[71]:
```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['RandomUnderSampler',
                                        RandomUnderSampler(random_state=300,
sampling_strategy='majority')],
                                       ['KNN', KNeighborsClassifier()]]),
             param_grid={'KNN__n_neighbors': [3, 4, 5, 6]},
             return_train_score=True, scoring='precision')
```

[72]:
```
y1_predict = knn_grid1.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[72]: 0.06285521501051047

2. Over-sampling the data

[61]:
```
# Pipeline with random over sampling and logistic regression for transformed␣
 ↪dataset

ros_log_pipeline1 = imbpipeline(steps = [['RandomOverSampler',␣
 ↪RandomOverSampler(sampling_strategy = 'minority',random_state = 300)],
                                        ['LogisticRegression',␣
 ↪LogisticRegression(random_state = 300)]])
```

[62]:
```
log_param = {
    'LogisticRegression__C': [1.0, 0.1, 0.01, 0.001]
}
log_grid = GridSearchCV(ros_log_pipeline1, param_grid = log_param, cv = skf,␣
 ↪scoring = 'precision', return_train_score = True)
```

[63]:
```
log_grid.fit(X1_train, y1_train)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
```

```
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
```

```
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

[63]:
```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['RandomOverSampler',
                                        RandomOverSampler(random_state=300,
sampling_strategy='minority')],
                                       ['LogisticRegression',
                                        LogisticRegression(random_state=300)]]),
             param_grid={'LogisticRegression__C': [1.0, 0.1, 0.01, 0.001]},
             return_train_score=True, scoring='precision')
```

[64]:
```
y1_predict = log_grid.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[64]: 0.03853340384276397

[65]:
```
# Pipeline with random over sampling and decision tree for transformed dataset

ros_tree_pipeline1 = imbpipeline(steps = [['RandomOverSampler',␣
 ↪RandomOverSampler(sampling_strategy = 'minority',random_state = 300)],
                                      ['DecisionTree', tree.
 ↪DecisionTreeClassifier(random_state = 300)]])
```

[66]:
```
tree_param1 = {
    'DecisionTree__criterion': ['gini', 'entropy'],
    'DecisionTree__max_depth': [5, 10, 20, 25]
}
tree_grid1 = GridSearchCV(ros_tree_pipeline1, param_grid = tree_param1, cv =␣
 ↪skf, scoring = 'precision', return_train_score = True)
```

[67]:
```
tree_grid1.fit(X1_train, y1_train)
```

[67]:
```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['RandomOverSampler',
                                        RandomOverSampler(random_state=300,
sampling_strategy='minority')],
                                       ['DecisionTree',
DecisionTreeClassifier(random_state=300)]]),
             param_grid={'DecisionTree__criterion': ['gini', 'entropy'],
                         'DecisionTree__max_depth': [5, 10, 20, 25]},
             return_train_score=True, scoring='precision')
```

[68]:
```
y1_predict = tree_grid1.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[68]:
```
0.3212688652878703
```

[ ]:
```
# Pipeline with random over sampling and KNN for transformed dataset

ros_knn_pipeline1 = imbpipeline(steps = [['RandomOverSampler',␣
 ↪RandomOverSampler(sampling_strategy = 'minority',random_state = 300)],
                                        ['KNN', KNeighborsClassifier()]])
```

[ ]:
```
knn_param1 = {
    'KNN__n_neighbors': [3, 4, 5, 6]
}
knn_grid1 = GridSearchCV(ros_knn_pipeline1, param_grid = knn_param1, cv = skf,␣
 ↪scoring = 'precision', return_train_score = True)
```

[ ]:
```
knn_grid1.fit(X1_train, y1_train)
```

3. SMOTE

[73]:
```
# Pipeline with SMOTE and logistic regression for transformed dataset

smote_log_pipeline1 = imbpipeline(steps = [['SMOTE', SMOTE(random_state = 300)],
                                        ['LogisticRegression',␣
 ↪LogisticRegression(random_state = 300)]])
```

[74]:
```
log_param = {
    'LogisticRegression__C': [1.0, 0.1, 0.01, 0.001]
}
log_grid = GridSearchCV(smote_log_pipeline1, param_grid = log_param, cv = skf,␣
 ↪scoring = 'precision', return_train_score = True)
```

[75]:
```
log_grid.fit(X1_train, y1_train)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
```

```
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
```

```
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458:
ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

[75]:
```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['SMOTE', SMOTE(random_state=300)],
                                       ['LogisticRegression',
                                        LogisticRegression(random_state=300)]]),
             param_grid={'LogisticRegression__C': [1.0, 0.1, 0.01, 0.001]},
             return_train_score=True, scoring='precision')
```

[76]:
```
y1_predict = log_grid.predict(X1_test)
precision_score(y1_test, y1_predict)
```

[76]:
```
0.20663291662361813
```

[77]:
```
# Pipeline with SMOTE and decision tree for transformed dataset

smote_tree_pipeline1 = imbpipeline(steps = [['SMOTE', SMOTE(random_state =␣
 ↪300)],
                                            ['DecisionTree', tree.
 ↪DecisionTreeClassifier(random_state = 300)]])
```

[78]:
```
tree_param1 = {
    'DecisionTree__criterion': ['gini', 'entropy'],
    'DecisionTree__max_depth': [5, 10, 20, 25]
}
tree_grid1 = GridSearchCV(smote_tree_pipeline1, param_grid = tree_param1, cv =␣
 ↪skf, scoring = 'precision', return_train_score = True)
```

[79]:
```
tree_grid1.fit(X1_train, y1_train)
```

```
[79]: GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=300, shuffle=True),
             estimator=Pipeline(steps=[['SMOTE', SMOTE(random_state=300)],
                                       ['DecisionTree',
DecisionTreeClassifier(random_state=300)]]),
             param_grid={'DecisionTree__criterion': ['gini', 'entropy'],
                         'DecisionTree__max_depth': [5, 10, 20, 25]},
             return_train_score=True, scoring='precision')
```

```
[80]: y1_predict = tree_grid1.predict(X1_test)
precision_score(y1_test, y1_predict)
```

```
[80]: 0.31074420896993593
```

```
[ ]: # Pipeline with SMOTE and KNN for transformed dataset

smote_knn_pipeline1 = imbpipeline(steps = [['SMOTE', SMOTE(random_state = 300)],
                                           ['KNN', KNeighborsClassifier()]])
```

```
[ ]: knn_param1 = {
    'KNN__n_neighbors': [3, 4, 5, 6]
}
knn_grid1 = GridSearchCV(smote_knn_pipeline1, param_grid = knn_param1, cv =␣
 ↪skf, scoring = 'precision', return_train_score = True)
```

```
[ ]: knn_grid1.fit(X1_train, y1_train)
```

```
[ ]: y1_predict = knn_grid1.predict(X1_test)
precision_score(y1_test, y1_predict)
```

Analyzing some of the models that have been runned and using a precison score to get a preview of the models' performance, it is observed that each of the models from the 3 resampling techniques performed relatively poorly. Essentially, the purpose of the precison score is to measure the proportion of the observations that were predicted as fraudluent that is actually fraudulent and vice versa in which a higher value closer to 1 is desirable. Knowing this, we can see that out of all the resampling techniques, although still fairly low, SMOTE stood out to be a better performer in comparison to undersampling and oversampling (logistic regression = 0.20, decision tree = 0.31) with decision tree being the better performing machine learning algorithm. As for KNN, it was only runned for the under sampling segement since it had a longer run time (approximately 1 hour for the algorithm to run). In fact, this issue applies for almost all the models when fitting in the training set. With this in mind, the next steps may include adjusting the parameters of the models and possibly look into other algorithms that may have shorter run times.