subramanian540_Project_Milestone2

April 23, 2023

1 DSC 540 - Data Preparation

2 Week5 and Week 6

3 Project Milestone 2

4 Cleaning/Formatting Flat File Source

Perform at least 5 data transformation and/or cleansing steps to your flat file data. The flat file that I have chosen for my term project is a Credit Card Transactions Fraud Detection Dataset.

Because this is a package in R, the data is pretty clean to begin with. The 5 data transformations that I will do are as follows:

Check for missing values in any of the columns that will be kept in the final data set.

Delete the transaciton number as 'trans_num' column.

Delete the Merchant Latitude as 'Merch_lat' column.

Delete the Merchant Longitude as 'Merch_long' column.

These columns will be deleted since at this time it they do not provide any information pertinent to the term project.

Convert time as 'unix_time' to a readable format

The Unix_time transformation, the 'unix_time' column will contain datetime objects that can be used for various types of analyses or visualizations that require time-based data

Add a column heading of 'row_id' to the first column.

This will be done because the first column does not have a header.

Convert 'amt' to a float with two decimal places

The 'amt' transformation column will contain floating-point values with two decimal places, which can be more appropriate for financial analyses or visualizations that require the display of currency amounts.

Identify outliers and bad data

The IQR is calculated by subtracting the first quartile (Q1) from the third quartile (Q3), which defines the middle 50% of the data. Outliers are then identified as any values that fall outside of the range of 1.5 times the IQR below Q1 or above Q3.Identifies any values in the 'amt' column that fall outside the range of 1.5 times the IQR below Q1 or above Q3 using boolean indexing.These are significantly different from other values in the dataset and may need to be examined more closely to determine if they are valid data points or errors that need to be corrected

Check for any duplcate values in any of the columns that will be kept in the final data set.

Drop the duplicates rows that will be kept in the final data set

Ethical Implications: The ethical implications of data transformation are critical to consider, particularly when dealing with sensitive data such as credit card fraud data. The process of data transformation involves handling missing values, removing irrelevant columns, converting data formats, identifying outliers, and removing duplicates.The data includes various fields such as transaction amount, transaction time, merchant ID, and customer information. However, due to the sensitive nature of this data, there are several ethical considerations to keep in mind during the data transformation process. It's essential to ensure that the data is only accessed by authorized personnel and protected against unauthorized access or theft. The transformed data should be used solely for legitimate purposes such as fraud detection, prevention, or investigation, and not for any illegal or unethical activities. Finally, it's essential to ensure that the data is deleted securely once it's no longer needed to prevent any further exposure of sensitive information.

[43]:

```
#Load the Necessary Libraries

import requests as r
import pandas as pd
import xlrd
from bs4 import BeautifulSoup
import numpy as np
import matplotlib.pyplot as plt
```

[44]:

```
# Load the FraudTest.csv file

fraud_data = pd.read_csv('fraudTest.csv', sep=",")
fraud_data.head()
```

[44]:

```
   Unnamed: 0 trans_date_trans_time        cc_num  \
0           0       6/21/2020 12:14  2.291160e+15
1           1       6/21/2020 12:14  3.573030e+15
2           2       6/21/2020 12:14  3.598220e+15
3           3       6/21/2020 12:15  3.591920e+15
4           4       6/21/2020 12:15  3.526830e+15

                               merchant        category   amt   first  \
0                 fraud_Kirlin and Sons   personal_care  2.86    Jeff
1                  fraud_Sporer-Keebler   personal_care 29.84  Joanne
2  fraud_Swaniawski, Nitzsche and Welch  health_fitness  41.28  Ashley
3                     fraud_Haley Group        misc_pos  60.05   Brian
4                 fraud_Johnston-Casper          travel   3.19  Nathan

       last gender                       street  …      lat      long  \
0   Elliott      M           351 Darlene Green  …  33.9659  -80.9355
1  Williams      F            3638 Marsh Union  …  40.3207 -110.4360
2     Lopez      F        9333 Valentine Point  …  40.6729  -73.5365
3  Williams      M  32941 Krystal Mill Apt. 552  …  28.5697  -80.8191
4    Massey      M     5783 Evan Roads Apt. 465  …  44.2529  -85.0170

   city_pop                        job        dob  \
0    333497     Mechanical engineer  3/19/1968
1       302  Sales professional, IT  1/17/1990
2     34496       Librarian, public  10/21/1970
3     54767            Set designer  7/25/1987
4      1126      Furniture designer    7/6/1955

                          trans_num   unix_time  merch_lat  merch_long  \
0  2da90c7d74bd46a0caf3777415b3ebd3  1371816865  33.986391  -81.200714
1  324cc204407e99f51b0d6ca0055005e7  1371816873  39.450498 -109.960431
2  c81755dbbbea9d5c77f094348a7579be  1371816893  40.495810  -74.196111
3  2159175b9efe66dc301f149d3d5abf8c  1371816915  28.812398  -80.883061
4  57ff021bd3f328f8738bb535c302a31b  1371816917  44.959148  -85.884734

   is_fraud
0         0
1         0
2         0
3         0
4         0

[5 rows x 23 columns]
```

[45]:

```python
# Transformation 1: Check for missing values in any of the columns that will be␣
 ↪kept in the final data set.

for c in fraud_data.columns:
    miss = fraud_data[c].isnull().sum()
    if miss>0:
        print("{} has {} missing value(s).".format(c,miss))
    else:
        print("{} has no missing values.".format(c))
```

```
Unnamed: 0 has no missing values.
trans_date_trans_time has 16 missing value(s).
cc_num has no missing values.
merchant has no missing values.
category has no missing values.
amt has no missing values.
first has no missing values.
last has no missing values.
gender has no missing values.
```

```
street has no missing values.
city has no missing values.
state has no missing values.
zip has no missing values.
lat has 14 missing value(s).
long has no missing values.
city_pop has no missing values.
job has no missing values.
dob has no missing values.
trans_num has no missing values.
unix_time has no missing values.
merch_lat has 13 missing value(s).
merch_long has no missing values.
is_fraud has no missing values.
```

[46]:
```
# Transformation 2. Delete the 'trans_num' column.

del fraud_data['trans_num']
fraud_data
```

[46]:
```
        Unnamed: 0 trans_date_trans_time        cc_num  \
0                0       6/21/2020 12:14  2.291160e+15
1                1       6/21/2020 12:14  3.573030e+15
2                2       6/21/2020 12:14  3.598220e+15
3                3       6/21/2020 12:15  3.591920e+15
4                4       6/21/2020 12:15  3.526830e+15
…             …                   …            …
555741      555714      12/31/2020 23:59  3.056060e+13
555742      555715      12/31/2020 23:59  3.556610e+15
555743      555716      12/31/2020 23:59  6.011720e+15
555744      555717      12/31/2020 23:59  4.079770e+12
555745      555718      12/31/2020 23:59  4.170690e+15

                                    merchant        category     amt    first  \
0                       fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                        fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                           fraud_Haley Group        misc_pos   60.05    Brian
4                       fraud_Johnston-Casper          travel    3.19   Nathan
…                                         …             …      …        …
555741                  fraud_Reilly and Sons  health_fitness   43.77  Michael
555742                   fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555743                       fraud_Rau-Robel       kids_pets   86.88      Ann
555744                  fraud_Breitenberg LLC          travel    7.99     Eric
555745                      fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street  …    zip      lat  \
0        Elliott      M            351 Darlene Green  …  29209  33.9659
1       Williams      F             3638 Marsh Union  …  84002  40.3207
2          Lopez      F         9333 Valentine Point  …  11710  40.6729
3       Williams      M  32941 Krystal Mill Apt. 552  …  32780  28.5697
4         Massey      M     5783 Evan Roads Apt. 465  …  49632  44.2529
…            …   …                           …  …   …       …
555741     Olson      M          558 Michael Estates  …  63453  40.4931
555742   Vasquez      M          572 Davis Mountains  …  77566  29.0393
555743    Lawson      F   144 Evans Islands Apt. 683  …  99323  46.1966
555744   Preston      M   7020 Doyle Stream Apt. 951  …  83643  44.6255
555745      Frey      M     830 Myers Plaza Apt. 384  …  73034  35.6665

            long  city_pop                     job         dob   unix_time  \
0       -80.9355    333497      Mechanical engineer   3/19/1968  1371816865
1      -110.4360       302   Sales professional, IT   1/17/1990  1371816873
2       -73.5365     34496       Librarian, public  10/21/1970  1371816893
3       -80.8191     54767             Set designer   7/25/1987  1371816915
4       -85.0170      1126       Furniture designer    7/6/1955  1371816917
…            …         …                        …          …          …
555741  -91.8912       519             Town planner   2/13/1966  1388534347
555742  -95.4401     28739           Futures trader  12/27/1999  1388534349
555743 -118.9017      3684                 Musician  11/29/1981  1388534355
555744 -116.4493       129             Cartographer  12/15/1965  1388534364
555745  -97.4798    116001              Media buyer   5/10/1993  1388534374

          merch_lat  merch_long  is_fraud
0         33.986391  -81.200714         0
1         39.450498 -109.960431         0
2         40.495810  -74.196111         0
3         28.812398  -80.883061         0
4         44.959148  -85.884734         0
…               …          …        …
555741    39.946837  -91.333331         0
555742    29.661049  -96.186633         0
555743    46.658340 -119.715054         0
555744    44.470525 -117.080888         0
555745    36.210097  -97.036372         0

[555746 rows x 22 columns]
```

[47]:

```
# Transformation 3. Delete the 'merch_lat ' column.

del fraud_data['merch_lat']
fraud_data
```

[47]:

```
        Unnamed: 0 trans_date_trans_time        cc_num  \
0                0       6/21/2020 12:14  2.291160e+15
```

```
1                 1       6/21/2020 12:14  3.573030e+15
2                 2       6/21/2020 12:14  3.598220e+15
3                 3       6/21/2020 12:15  3.591920e+15
4                 4       6/21/2020 12:15  3.526830e+15
…            …                   …           …
555741      555714      12/31/2020 23:59  3.056060e+13
555742      555715      12/31/2020 23:59  3.556610e+15
555743      555716      12/31/2020 23:59  6.011720e+15
555744      555717      12/31/2020 23:59  4.079770e+12
555745      555718      12/31/2020 23:59  4.170690e+15

                                    merchant        category     amt    first  \
0                      fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                          fraud_Haley Group        misc_pos   60.05    Brian
4                      fraud_Johnston-Casper          travel    3.19   Nathan
…                                      …             …     …       …
555741                 fraud_Reilly and Sons  health_fitness   43.77  Michael
555742                  fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555743                       fraud_Rau-Robel       kids_pets   86.88      Ann
555744                 fraud_Breitenberg LLC          travel    7.99     Eric
555745                     fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street  … state    zip  \
0        Elliott      M            351 Darlene Green  …    SC  29209
1       Williams      F             3638 Marsh Union  …    UT  84002
2          Lopez      F         9333 Valentine Point  …    NY  11710
3       Williams      M  32941 Krystal Mill Apt. 552  …    FL  32780
4         Massey      M     5783 Evan Roads Apt. 465  …    MI  49632
…          …    …                          …  …   …    …
555741     Olson      M          558 Michael Estates  …    MO  63453
555742   Vasquez      M          572 Davis Mountains  …    TX  77566
555743    Lawson      F   144 Evans Islands Apt. 683  …    WA  99323
555744   Preston      M   7020 Doyle Stream Apt. 951  …    ID  83643
555745      Frey      M     830 Myers Plaza Apt. 384  …    OK  73034

            lat      long  city_pop                     job         dob  \
0       33.9659  -80.9355    333497     Mechanical engineer   3/19/1968
1       40.3207 -110.4360       302  Sales professional, IT   1/17/1990
2       40.6729  -73.5365     34496       Librarian, public  10/21/1970
3       28.5697  -80.8191     54767            Set designer   7/25/1987
4       44.2529  -85.0170      1126      Furniture designer    7/6/1955
…         …       …       …                     …         …
555741  40.4931  -91.8912       519            Town planner   2/13/1966
555742  29.0393  -95.4401     28739          Futures trader  12/27/1999
555743  46.1966 -118.9017      3684                Musician  11/29/1981
```

```
555744  44.6255 -116.4493      129          Cartographer  12/15/1965
555745  35.6665  -97.4798   116001           Media buyer   5/10/1993

         unix_time  merch_long  is_fraud
0       1371816865  -81.200714         0
1       1371816873 -109.960431         0
2       1371816893  -74.196111         0
3       1371816915  -80.883061         0
4       1371816917  -85.884734         0
…              …         …       …
555741  1388534347  -91.333331         0
555742  1388534349  -96.186633         0
555743  1388534355 -119.715054         0
555744  1388534364 -117.080888         0
555745  1388534374  -97.036372         0

[555746 rows x 21 columns]
```

[48]:

```
# Transformation 4. Delete the 'merch_long' column.

del fraud_data['merch_long']
fraud_data
```

[48]:

```
        Unnamed: 0 trans_date_trans_time        cc_num  \
0                0       6/21/2020 12:14  2.291160e+15
1                1       6/21/2020 12:14  3.573030e+15
2                2       6/21/2020 12:14  3.598220e+15
3                3       6/21/2020 12:15  3.591920e+15
4                4       6/21/2020 12:15  3.526830e+15
…              …                    …           …
555741      555714      12/31/2020 23:59  3.056060e+13
555742      555715      12/31/2020 23:59  3.556610e+15
555743      555716      12/31/2020 23:59  6.011720e+15
555744      555717      12/31/2020 23:59  4.079770e+12
555745      555718      12/31/2020 23:59  4.170690e+15

                                   merchant        category     amt     first  \
0                      fraud_Kirlin and Sons   personal_care    2.86      Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84    Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28    Ashley
3                          fraud_Haley Group        misc_pos   60.05     Brian
4                      fraud_Johnston-Casper          travel    3.19    Nathan
…                                      …             …    …       …
555741                 fraud_Reilly and Sons  health_fitness   43.77   Michael
555742                  fraud_Hoppe-Parisian       kids_pets  111.84      Jose
555743                      fraud_Rau-Robel       kids_pets   86.88       Ann
555744                 fraud_Breitenberg LLC          travel    7.99      Eric
```

```
555745                     fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street          city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      F             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                          …           …   …
555741     Olson      M          558 Michael Estates         Luray    MO
555742   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555743    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555744   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555745      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat      long  city_pop                     job  \
0       29209  33.9659  -80.9355    333497     Mechanical engineer
1       84002  40.3207 -110.4360       302  Sales professional, IT
2       11710  40.6729  -73.5365     34496       Librarian, public
3       32780  28.5697  -80.8191     54767            Set designer
4       49632  44.2529  -85.0170      1126      Furniture designer
…       …      …       …       …                      …
555741  63453  40.4931  -91.8912       519            Town planner
555742  77566  29.0393  -95.4401     28739          Futures trader
555743  99323  46.1966 -118.9017      3684                Musician
555744  83643  44.6255 -116.4493       129            Cartographer
555745  73034  35.6665  -97.4798    116001             Media buyer

               dob   unix_time  is_fraud
0        3/19/1968  1371816865         0
1        1/17/1990  1371816873         0
2       10/21/1970  1371816893         0
3        7/25/1987  1371816915         0
4         7/6/1955  1371816917         0
…            …         …       …
555741   2/13/1966  1388534347         0
555742  12/27/1999  1388534349         0
555743  11/29/1981  1388534355         0
555744  12/15/1965  1388534364         0
555745   5/10/1993  1388534374         0

[555746 rows x 20 columns]
```

[49]:

```
# Transformation 5.Convert time to a readable format

fraud_data['unix_time'] = pd.to_datetime(fraud_data['unix_time'], unit='s')
fraud_data
```

```
[49]:         Unnamed: 0 trans_date_trans_time        cc_num  \
0                0       6/21/2020 12:14  2.291160e+15
1                1       6/21/2020 12:14  3.573030e+15
2                2       6/21/2020 12:14  3.598220e+15
3                3       6/21/2020 12:15  3.591920e+15
4                4       6/21/2020 12:15  3.526830e+15
…            …                   …           …
555741      555714      12/31/2020 23:59  3.056060e+13
555742      555715      12/31/2020 23:59  3.556610e+15
555743      555716      12/31/2020 23:59  6.011720e+15
555744      555717      12/31/2020 23:59  4.079770e+12
555745      555718      12/31/2020 23:59  4.170690e+15

                                    merchant        category     amt    first  \
0                      fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                          fraud_Haley Group        misc_pos   60.05    Brian
4                      fraud_Johnston-Casper          travel    3.19   Nathan
…                                      …             …     …      …
555741                 fraud_Reilly and Sons  health_fitness   43.77  Michael
555742                  fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555743                       fraud_Rau-Robel       kids_pets   86.88      Ann
555744                 fraud_Breitenberg LLC          travel    7.99     Eric
555745                     fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street          city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      F             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                          …           …   …
555741     Olson      M          558 Michael Estates         Luray    MO
555742   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555743    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555744   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555745      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat       long  city_pop                     job  \
0       29209  33.9659   -80.9355    333497     Mechanical engineer
1       84002  40.3207  -110.4360       302  Sales professional, IT
2       11710  40.6729   -73.5365     34496       Librarian, public
3       32780  28.5697   -80.8191     54767            Set designer
4       49632  44.2529   -85.0170      1126      Furniture designer
…       …      …       …       …                     …
555741  63453  40.4931   -91.8912       519            Town planner
```

```
555742  77566  29.0393  -95.4401     28739          Futures trader
555743  99323  46.1966 -118.9017      3684                Musician
555744  83643  44.6255 -116.4493       129            Cartographer
555745  73034  35.6665  -97.4798    116001             Media buyer

               dob           unix_time  is_fraud
0        3/19/1968 2013-06-21 12:14:25         0
1        1/17/1990 2013-06-21 12:14:33         0
2       10/21/1970 2013-06-21 12:14:53         0
3        7/25/1987 2013-06-21 12:15:15         0
4         7/6/1955 2013-06-21 12:15:17         0
…            …                 …       …
555741   2/13/1966 2013-12-31 23:59:07         0
555742  12/27/1999 2013-12-31 23:59:09         0
555743  11/29/1981 2013-12-31 23:59:15         0
555744  12/15/1965 2013-12-31 23:59:24         0
555745   5/10/1993 2013-12-31 23:59:34         0

[555746 rows x 20 columns]
```

[50]:

```
# Transformation 6. Add a column heading of 'row_id' to the first column.

fraud_data.rename(columns = {'Unnamed: 0' : 'row_id'}, inplace=True)
fraud_data
```

[50]:

```
        row_id trans_date_trans_time        cc_num  \
0            0       6/21/2020 12:14  2.291160e+15
1            1       6/21/2020 12:14  3.573030e+15
2            2       6/21/2020 12:14  3.598220e+15
3            3       6/21/2020 12:15  3.591920e+15
4            4       6/21/2020 12:15  3.526830e+15
…        …                   …           …
555741  555714      12/31/2020 23:59  3.056060e+13
555742  555715      12/31/2020 23:59  3.556610e+15
555743  555716      12/31/2020 23:59  6.011720e+15
555744  555717      12/31/2020 23:59  4.079770e+12
555745  555718      12/31/2020 23:59  4.170690e+15

                                    merchant        category     amt     first  \
0                      fraud_Kirlin and Sons   personal_care    2.86      Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84    Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28    Ashley
3                          fraud_Haley Group        misc_pos   60.05     Brian
4                      fraud_Johnston-Casper          travel    3.19    Nathan
…                                      …             …     …      …
555741                 fraud_Reilly and Sons  health_fitness   43.77   Michael
555742                  fraud_Hoppe-Parisian       kids_pets  111.84      Jose
```

```
555743                       fraud_Rau-Robel       kids_pets   86.88      Ann
555744                 fraud_Breitenberg LLC          travel    7.99     Eric
555745                   fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street          city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      F             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                          …           …   …
555741     Olson      M          558 Michael Estates         Luray    MO
555742   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555743    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555744   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555745      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat      long  city_pop                     job  \
0       29209  33.9659  -80.9355    333497     Mechanical engineer
1       84002  40.3207 -110.4360       302  Sales professional, IT
2       11710  40.6729  -73.5365     34496       Librarian, public
3       32780  28.5697  -80.8191     54767            Set designer
4       49632  44.2529  -85.0170      1126      Furniture designer
…       …      …       …       …                     …
555741  63453  40.4931  -91.8912       519            Town planner
555742  77566  29.0393  -95.4401     28739          Futures trader
555743  99323  46.1966 -118.9017      3684                Musician
555744  83643  44.6255 -116.4493       129            Cartographer
555745  73034  35.6665  -97.4798    116001             Media buyer

               dob           unix_time  is_fraud
0        3/19/1968 2013-06-21 12:14:25        0
1        1/17/1990 2013-06-21 12:14:33        0
2       10/21/1970 2013-06-21 12:14:53        0
3        7/25/1987 2013-06-21 12:15:15        0
4         7/6/1955 2013-06-21 12:15:17        0
…            …                 …       …
555741   2/13/1966 2013-12-31 23:59:07        0
555742  12/27/1999 2013-12-31 23:59:09        0
555743  11/29/1981 2013-12-31 23:59:15        0
555744  12/15/1965 2013-12-31 23:59:24        0
555745   5/10/1993 2013-12-31 23:59:34        0

[555746 rows x 20 columns]
```

[51]:

```
# Transformation 7. Convert amount to a float with two decimal places
```

```python
fraud_data['amt'] = fraud_data['amt'].round(2)
fraud_data
```

[51]:

```
        row_id trans_date_trans_time        cc_num  \
0            0       6/21/2020 12:14  2.291160e+15
1            1       6/21/2020 12:14  3.573030e+15
2            2       6/21/2020 12:14  3.598220e+15
3            3       6/21/2020 12:15  3.591920e+15
4            4       6/21/2020 12:15  3.526830e+15
…        …                   …           …
555741  555714      12/31/2020 23:59  3.056060e+13
555742  555715      12/31/2020 23:59  3.556610e+15
555743  555716      12/31/2020 23:59  6.011720e+15
555744  555717      12/31/2020 23:59  4.079770e+12
555745  555718      12/31/2020 23:59  4.170690e+15

                                    merchant        category     amt    first  \
0                      fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                          fraud_Haley Group        misc_pos   60.05    Brian
4                      fraud_Johnston-Casper          travel    3.19   Nathan
…                                      …             …     …      …
555741                 fraud_Reilly and Sons  health_fitness   43.77  Michael
555742                  fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555743                       fraud_Rau-Robel       kids_pets   86.88      Ann
555744                 fraud_Breitenberg LLC          travel    7.99     Eric
555745                     fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street          city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      M             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                          …           …   …
555741     Olson      M          558 Michael Estates         Luray    MO
555742   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555743    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555744   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555745      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat      long  city_pop                     job  \
0       29209  33.9659  -80.9355    333497     Mechanical engineer
1       84002  40.3207 -110.4360       302  Sales professional, IT
2       11710  40.6729  -73.5365     34496       Librarian, public
3       32780  28.5697  -80.8191     54767            Set designer
```

```
4        49632  44.2529  -85.0170      1126      Furniture designer
…        …      …        …        …                        …
555741   63453  40.4931  -91.8912       519            Town planner
555742   77566  29.0393  -95.4401     28739          Futures trader
555743   99323  46.1966 -118.9017      3684                Musician
555744   83643  44.6255 -116.4493       129            Cartographer
555745   73034  35.6665  -97.4798    116001             Media buyer

              dob           unix_time  is_fraud
0       3/19/1968 2013-06-21 12:14:25         0
1       1/17/1990 2013-06-21 12:14:33         0
2      10/21/1970 2013-06-21 12:14:53         0
3       7/25/1987 2013-06-21 12:15:15         0
4        7/6/1955 2013-06-21 12:15:17         0
…             …                   …        …
555741  2/13/1966 2013-12-31 23:59:07         0
555742 12/27/1999 2013-12-31 23:59:09         0
555743 11/29/1981 2013-12-31 23:59:15         0
555744 12/15/1965 2013-12-31 23:59:24         0
555745  5/10/1993 2013-12-31 23:59:34         0

[555746 rows x 20 columns]
```

[52]:

```python
# Transformation 8 .identify outliers and bad data

Q1 = fraud_data['amt'].quantile(0.25)
Q3 = fraud_data['amt'].quantile(0.75)
IQR = Q3 - Q1

outliers = fraud_data[(fraud_data['amt'] < (Q1 - 1.5 * IQR)) |␣
 ↪(fraud_data['amt'] > (Q3 + 1.5 * IQR))]

print("Outliers:\n", outliers)
```

```
Outliers:
         row_id trans_date_trans_time        cc_num                  merchant  \
33          33       6/21/2020 12:26  1.800360e+14          fraud_Nienow PLC
100        100       6/21/2020 12:45  6.592070e+15       fraud_Greenholt Ltd
133        133       6/21/2020 12:55  4.683640e+12      fraud_Gislason Group
167        167       6/21/2020 13:08  6.011110e+15          fraud_Schumm PLC
245        245       6/21/2020 13:36  4.476840e+12      fraud_Heathcote LLC
…          …                    …           …                        …
555526  555526      12/31/2020 22:51  4.810840e+18           fraud_Swift PLC
555625  555625      12/31/2020 23:24  6.540980e+15    fraud_Weimann-Lockman
555629  555629      12/31/2020 23:26  2.248740e+15    fraud_Crooks and Sons
555637  555637      12/31/2020 23:30  6.011110e+15 fraud_Gleason-Macejkovic
555647  555647      12/31/2020 23:32  6.011440e+15           fraud_Auer-West
```

```
              category      amt      first      last gender  \
33       entertainment   210.36  Mackenzie   Salazar      F   
100     health_fitness   242.35     Amanda    Molina      F   
133             travel   558.03     Daniel      Boyd      M   
167       shopping_net  1199.45    Rebecca  Erickson      F   
245       shopping_net   236.15     Steven   Walters      M   
…                …      …        …       …    …               
555526       kids_pets   290.11    Carolyn     Perez      F   
555625       kids_pets   255.42      Bryan    Torres      M   
555629   personal_care   302.79      Jacob     Weber      M   
555637    shopping_net  1164.37    Rebecca  Erickson      F   
555647    shopping_net   410.05    Allison     Allen      F   

                              street          city state    zip      lat  \
33                  982 Melissa Lock        Bagley    WI  53801  42.9207   
100     8425 Daniel Knolls Suite 288  Philadelphia    PA  19154  40.0897   
133             8925 Nicholas Points          Egan    LA  70531  30.2510   
167        594 Berry Lights Apt. 392    Wilmington    NC  28405  34.2651   
245       3206 Hall Divide Suite 282     Woodville    AL  35776  34.6689   
…                              …           …   …    …      …             
555526     433 Blake Roads Suite 967       Wheaton    MO  64874  36.7651   
555625    152 James Centers Apt. 768       Detroit    MI  48221  42.4260   
555629      29156 Mark Park Apt. 108         Utica    KS  67584  38.6411   
555637     594 Berry Lights Apt. 392    Wilmington    NC  28405  34.2651   
555647           40624 Rebecca Spurs       De Witt    AR  72042  34.2853   

            long  city_pop                                    job         dob  \
33      -91.0685       878                           Risk analyst  11/20/1974   
100     -74.9781   1526206              Commercial horticulturist   5/23/1972   
133     -92.5002      1261                    Broadcast presenter    7/1/1972   
167     -77.8670    186140   English as a second language teacher    2/8/1983   
245     -86.2296      3395                  Editor, commissioning   1/21/1979   
…          …       …                                    …         …              
555526  -94.0492       760                     Production manager   8/31/1985   
555625  -83.1500    673342                         Retail manager   6/19/1967   
555629 -100.1380       269  Product/process development scientist  11/11/1962   
555637  -77.8670    186140   English as a second language teacher    2/8/1983   
555647  -91.3336      5161                    Electrical engineer    4/8/1993   

                  unix_time  is_fraud  
33      2013-06-21 12:26:23         0  
100     2013-06-21 12:45:48         0  
133     2013-06-21 12:55:19         0  
167     2013-06-21 13:08:46         0  
245     2013-06-21 13:36:21         0  
…                     …       …        
555526  2013-12-31 22:51:48         0  
555625  2013-12-31 23:24:37         0  
```

```
555629 2013-12-31 23:26:42         0
555637 2013-12-31 23:30:29         0
555647 2013-12-31 23:32:24         0

[27778 rows x 20 columns]
```

[53]:

```
# Transformation 9. Identify any missing values

missing_values = fraud_data.isnull().sum().sum()
print("Missing Values:\n", missing_values)
```

```
Missing Values:
 30
```

[54]:

```
# Transformation 10. Identify any duplicate rows

duplicates = fraud_data[fraud_data.duplicated()]
print("Duplicates:\n", duplicates)
```

```
Duplicates:
         row_id trans_date_trans_time        cc_num  \
555735  555708      12/31/2020 23:56  2.131120e+14
555736  555709      12/31/2020 23:57  3.034470e+13
555737  555710      12/31/2020 23:57  3.524570e+15
555738  555711      12/31/2020 23:57  3.415460e+14
555739  555712      12/31/2020 23:58  5.018030e+11
555740  555713      12/31/2020 23:58  3.523840e+15
555741  555714      12/31/2020 23:59  3.056060e+13
555742  555715      12/31/2020 23:59  3.556610e+15
555743  555716      12/31/2020 23:59  6.011720e+15
555744  555717      12/31/2020 23:59  4.079770e+12
555745  555718      12/31/2020 23:59  4.170690e+15

                                     merchant        category     amt  \
555735   fraud_Baumbach, Hodkiewicz and Walsh    shopping_pos   25.49
555736  fraud_Larkin, Stracke and Greenfelder   entertainment   46.71
555737    fraud_Heathcote, Yost and Kertzmann    shopping_net   29.56
555738                   fraud_Schmidt-Larkin            home   12.68
555739       fraud_Pouros, Walker and Spencer       kids_pets   13.02
555740    fraud_Prosacco, Kreiger and Kovacek            home   17.00
555741                  fraud_Reilly and Sons  health_fitness   43.77
555742                   fraud_Hoppe-Parisian       kids_pets  111.84
555743                        fraud_Rau-Robel       kids_pets   86.88
555744                  fraud_Breitenberg LLC          travel    7.99
555745                      fraud_Dare-Marvin   entertainment   38.13

            first      last gender                         street  \
555735        Ana    Howell      F  4664 Sanchez Common Suite 930
555736  Christine   Johnson      F   8011 Chapman Tunnel Apt. 568
```

```
555737     Ashley   Cabrera      F   94225 Smith Springs Apt. 617
555738       Mark     Brown      M                8580 Moore Cove
555739     Robert    Flores      M   3277 Fields Meadows Apt. 790
555740      Grace  Williams      F    28812 Charles Mill Apt. 628
555741    Michael     Olson      M            558 Michael Estates
555742       Jose   Vasquez      M            572 Davis Mountains
555743        Ann    Lawson      F     144 Evans Islands Apt. 683
555744       Eric   Preston      M     7020 Doyle Stream Apt. 951
555745     Samuel      Frey      M       830 Myers Plaza Apt. 384

                      city state    zip      lat      long  city_pop  \
555735             Bradley    SC  29819  34.0326  -82.2027      1523
555736  Blairsden-Graeagle    CA  96103  39.8127 -120.6405      1725
555737          Vero Beach    FL  32960  27.6330  -80.4031    105638
555738               Wales    AK  99783  64.7556 -165.6723       145
555739           Greenview    CA  96037  41.5403 -122.9366       308
555740       Plantersville    AL  36758  32.6176  -86.9475      1412
555741               Luray    MO  63453  40.4931  -91.8912       519
555742        Lake Jackson    TX  77566  29.0393  -95.4401     28739
555743             Burbank    WA  99323  46.1966 -118.9017      3684
555744                Mesa    ID  83643  44.6255 -116.4493       129
555745              Edmond    OK  73034  35.6665  -97.4798    116001

                                                  job         dob  \
555735         Research scientist (physical sciences)    6/3/1984
555736  Chartered legal executive (England and Wales)   5/27/1967
555737                              Librarian, public    5/7/1986
555738                       Administrator, education   11/9/1939
555739                            Call centre manager   9/20/1958
555740                              Drilling engineer  11/20/1970
555741                                   Town planner   2/13/1966
555742                                 Futures trader  12/27/1999
555743                                       Musician  11/29/1981
555744                                   Cartographer  12/15/1965
555745                                    Media buyer   5/10/1993

                 unix_time  is_fraud
555735 2013-12-31 23:56:57         0
555736 2013-12-31 23:57:18         0
555737 2013-12-31 23:57:50         0
555738 2013-12-31 23:57:56         0
555739 2013-12-31 23:58:04         0
555740 2013-12-31 23:58:34         0
555741 2013-12-31 23:59:07         0
555742 2013-12-31 23:59:09         0
555743 2013-12-31 23:59:15         0
555744 2013-12-31 23:59:24         0
555745 2013-12-31 23:59:34         0
```

[55]:

```
# Transformation 11. drop the duplicates
fraud_data = fraud_data.drop_duplicates()
fraud_data
```

[55]:

```
        row_id trans_date_trans_time        cc_num  \
0            0       6/21/2020 12:14  2.291160e+15
1            1       6/21/2020 12:14  3.573030e+15
2            2       6/21/2020 12:14  3.598220e+15
3            3       6/21/2020 12:15  3.591920e+15
4            4       6/21/2020 12:15  3.526830e+15
…        …                   …           …
555730  555714                   NaN  3.056060e+13
555731  555715                   NaN  3.556610e+15
555732  555716                   NaN  6.011720e+15
555733  555717                   NaN  4.079770e+12
555734  555718                   NaN  4.170690e+15

                                    merchant        category     amt    first  \
0                      fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                       fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                          fraud_Haley Group        misc_pos   60.05    Brian
4                      fraud_Johnston-Casper          travel    3.19   Nathan
…                                      …             …     …      …
555730                 fraud_Reilly and Sons  health_fitness   43.77  Michael
555731                  fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555732                       fraud_Rau-Robel       kids_pets   86.88      Ann
555733                 fraud_Breitenberg LLC          travel    7.99     Eric
555734                     fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street          city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      M             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                          …           …   …
555730     Olson      M          558 Michael Estates         Luray    MO
555731   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555732    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555733   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555734      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat      long  city_pop                  job  \
0       29209  33.9659  -80.9355    333497     Mechanical engineer
1       84002  40.3207 -110.4360       302  Sales professional, IT
2       11710  40.6729  -73.5365     34496       Librarian, public
```

```
3        32780  28.5697  -80.8191     54767            Set designer
4        49632  44.2529  -85.0170      1126      Furniture designer
…        …      …        …        …                     …
555730   63453      NaN  -91.8912       519            Town planner
555731   77566      NaN  -95.4401     28739          Futures trader
555732   99323      NaN -118.9017      3684                Musician
555733   83643      NaN -116.4493       129            Cartographer
555734   73034      NaN  -97.4798    116001             Media buyer

               dob           unix_time  is_fraud
0        3/19/1968 2013-06-21 12:14:25         0
1        1/17/1990 2013-06-21 12:14:33         0
2       10/21/1970 2013-06-21 12:14:53         0
3        7/25/1987 2013-06-21 12:15:15         0
4         7/6/1955 2013-06-21 12:15:17         0
…              …                 …       …
555730   2/13/1966 2013-12-31 23:59:07         0
555731  12/27/1999 2013-12-31 23:59:09         0
555732  11/29/1981 2013-12-31 23:59:15         0
555733  12/15/1965 2013-12-31 23:59:24         0
555734   5/10/1993 2013-12-31 23:59:34         0

[555735 rows x 20 columns]
```

[56]:

```python
# # Transformation 12. Verify data accuracy and Check  for negative 'amount'␣
 ↪values

negative_amounts = fraud_data[fraud_data['amt'] < 0]
if not negative_amounts.empty:
    # Replace negative values with NaN
    fraud_data.loc[fraud_data['amt'] < 0, 'amt'] = np.nan
    print("Negative amounts found and replaced with NaN.")
else:
    print("No negative amounts found.")
```

```
No negative amounts found.
```

[57]:

```python
# After all Transformation and the final Fraud data

fraud_data
```

[57]:

```
        row_id trans_date_trans_time        cc_num  \
0            0       6/21/2020 12:14  2.291160e+15
1            1       6/21/2020 12:14  3.573030e+15
2            2       6/21/2020 12:14  3.598220e+15
3            3       6/21/2020 12:15  3.591920e+15
4            4       6/21/2020 12:15  3.526830e+15
…          …                     …            …
```

```
555730  555714                   NaN  3.056060e+13
555731  555715                   NaN  3.556610e+15
555732  555716                   NaN  6.011720e+15
555733  555717                   NaN  4.079770e+12
555734  555718                   NaN  4.170690e+15

                                    merchant        category     amt    first  \
0                  fraud_Kirlin and Sons   personal_care    2.86     Jeff
1                   fraud_Sporer-Keebler   personal_care   29.84   Joanne
2       fraud_Swaniawski, Nitzsche and Welch  health_fitness   41.28   Ashley
3                      fraud_Haley Group        misc_pos   60.05    Brian
4                  fraud_Johnston-Casper          travel    3.19   Nathan
…                                      …             …     …      …
555730             fraud_Reilly and Sons  health_fitness   43.77  Michael
555731              fraud_Hoppe-Parisian       kids_pets  111.84     Jose
555732                   fraud_Rau-Robel       kids_pets   86.88      Ann
555733             fraud_Breitenberg LLC          travel    7.99     Eric
555734                 fraud_Dare-Marvin   entertainment   38.13   Samuel

            last gender                       street         city state  \
0        Elliott      M            351 Darlene Green      Columbia    SC
1       Williams      F             3638 Marsh Union       Altonah    UT
2          Lopez      F         9333 Valentine Point      Bellmore    NY
3       Williams      M  32941 Krystal Mill Apt. 552    Titusville    FL
4         Massey      M     5783 Evan Roads Apt. 465      Falmouth    MI
…          …    …                           …           …   …
555730     Olson      M          558 Michael Estates         Luray    MO
555731   Vasquez      M          572 Davis Mountains  Lake Jackson    TX
555732    Lawson      F   144 Evans Islands Apt. 683       Burbank    WA
555733   Preston      M   7020 Doyle Stream Apt. 951          Mesa    ID
555734      Frey      M     830 Myers Plaza Apt. 384        Edmond    OK

          zip      lat      long  city_pop                     job  \
0       29209  33.9659  -80.9355    333497     Mechanical engineer
1       84002  40.3207 -110.4360       302  Sales professional, IT
2       11710  40.6729  -73.5365     34496       Librarian, public
3       32780  28.5697  -80.8191     54767            Set designer
4       49632  44.2529  -85.0170      1126      Furniture designer
…           …      …        …       …                       …
555730  63453      NaN  -91.8912       519            Town planner
555731  77566      NaN  -95.4401     28739          Futures trader
555732  99323      NaN -118.9017      3684                Musician
555733  83643      NaN -116.4493       129            Cartographer
555734  73034      NaN  -97.4798    116001             Media buyer

               dob           unix_time  is_fraud
0        3/19/1968 2013-06-21 12:14:25         0
```

```
1        1/17/1990 2013-06-21 12:14:33         0
2       10/21/1970 2013-06-21 12:14:53         0
3        7/25/1987 2013-06-21 12:15:15         0
4         7/6/1955 2013-06-21 12:15:17         0
…            …                 …       …
555730   2/13/1966 2013-12-31 23:59:07         0
555731  12/27/1999 2013-12-31 23:59:09         0
555732  11/29/1981 2013-12-31 23:59:15         0
555733  12/15/1965 2013-12-31 23:59:24         0
555734   5/10/1993 2013-12-31 23:59:34         0

[555735 rows x 20 columns]
```

```
[ ]:
```