Part 2: Importing Data

a. Import the Pandas library with the alias pd and Import the blood_transfusion.csv file (provided via Canvas).

b. What are the dimensions of this data (number of rows and columns)?

c. What are the data types of each column?

d. Are there any missing values?

e. Check out the first 10 rows? What are the Class values for the first 10 observations?

f. Check out the last 10 rows? What are the Class values for the last 10 observations?

g. Index for the 100th row and just the Monetary column. What is the value?

h. Index for just the Monetary column. What is the mean of this vector?

In [5]:
```
import pandas as pd
df = pd.read_csv('blood_transfusion.csv')
```

In [7]:
```
#b:
df.shape
```

Out[7]:
```
(748, 5)
```

In [9]:
```
#c:
df.dtypes
```

Out[9]:
```
Recency       int64
Frequency     int64
Monetary      int64
Time          int64
Class        object
dtype: object
```

In [11]:
```
#d
df.isna().sum()
```

Out[11]:
```
Recency      0
Frequency    0
Monetary     0
Time         0
Class        0
dtype: int64
```

In [13]:
```
#e & f
df['Class'].head(10), df['Class'].tail(10)
```

Out[13]:
```
(0        donated
 1        donated
 2        donated
 3        donated
 4    not donated
 5    not donated
 6        donated
 7    not donated
 8        donated
 9        donated
 Name: Class, dtype: object,
 738    not donated
 739    not donated
 740    not donated
 741    not donated
 742    not donated
 743    not donated
 744    not donated
 745    not donated
 746    not donated
 747    not donated
 Name: Class, dtype: object)
```

In [17]:
```
#g
df.loc[99]['Monetary']
```

Out[17]:
```
1750
```

In [19]:
```
#h
df['Monetary'].mean()
```

Out[19]:
```
1378.6764705882354
```

In [27]:
```
dfmonetary = df[df['Monetary'] > df['Monetary'].mean()]
dfmonetary.shape
```

Out[27]:
```
(267, 5)
```

a. Import the SomePlaceWeatherData.csv file (provided via Canvas). The csv file has

headers as the first row, so use read_csv('SomePlaceWeatherData.csv', header=0).

b. What are the dimensions of this data (number of rows and columns)?

c. Are there any missing values in this data?

d. Index for the 365th row. What is the date of this observation and what was the

temperature?

e. Use indexing to get the first 31 days in our data frame and the Temperature

column. Is the output a Series or a DataFrame.

f. Use the .describe() method on this output to compute various summary statistics

such as min, max, mean, median, etc.

In [31]:
```
df1 = pd.read_csv('SomePlaceWeatherData.csv', header=0)
df1
```

Out[31]:

	Date	Temperature	Visibility(km)	Humidity	WindSpeed(km/h)	WindBearing(degrees)
0	1/1/2006	46	16	0.89	28	149
1	1/2/2006	46	16	0.99	27	239
2	1/3/2006	38	12	0.95	21	353
3	1/4/2006	38	12	0.96	18	330
4	1/5/2006	40	10	1.00	21	345
...	...	...	...	...	...	...
4013	12/27/2016	38	10	0.95	15	225
4014	12/28/2016	38	10	0.95	15	221
4015	12/29/2016	38	10	0.95	15	215
4016	12/30/2016	37	10	0.95	15	211
4017	12/31/2016	37	10	0.95	14	208

4018 rows × 6 columns

In [33]:
```
#b
df1.shape
```

Out[33]:
```
(4018, 6)
```

In [35]:
```
#c
df1.isna().sum()
```

Out[35]:
```
Date                    0
Temperature             0
Visibility(km)          0
Humidity                0
WindSpeed(km/h)         0
WindBearing(degrees)    0
dtype: int64
```

In [57]:
```
#d
df1.loc[364]['Date'], df1.loc[364]['Temperature']
```

Out[57]:
```
('12/31/2006', 36)
```

In [69]:
```
df1['Date'] = pd.to_datetime(df1['Date'])
subset = df1[df1['Date'] < '02/01/2006']
subset['Temperature'].max()
```

Out[69]:
```
46
```

In [71]:
```
df1['Temperature'].max()
```

Out[71]:
```
104
```

In [43]:
```
#e This is a series
df1.loc[0:30]['Temperature']
```

Out[43]:
```
0     46
1     46
2     38
3     38
4     40
5     38
6     37
7     37
8     37
9     38
10    35
11    35
12    33
13    32
14    29
15    28
16    29
17    39
18    37
19    33
20    44
21    36
22    14
23    20
24    24
25    27
26    32
27    33
28    42
29    35
30    32
Name: Temperature, dtype: int64
```

In [126…
```
#f
df1describe = df1.loc[0:30]['Temperature']
df1describe.describe()
```

Out[126…
```
count    31.000000
mean     34.322581
std       6.982775
min      14.000000
25%      32.000000
50%      35.000000
75%      38.000000
max      46.000000
Name: Temperature, dtype: float64
```

a. Import the PDIPolice_Data_InitiativeCrime_Incidents.csv data (provided via Canvas). Data is taken from the City of Cincinnati Open Data Portal website (https://data.cincinnati-oh.gov/safety/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf), which you may need to read to place context in your answers.

b. What are the dimensions of this data (number of rows and columns)?

c. Are there any missing values in this data? If so, how many missing values are in each column? Which column has the most missing values?

d. Index for the DATE_REPORTED column and apply the .describe() method to it. Which date has the most observations in this data set?

e. Index for the 100th observation in this data set. What was the date of this crime? What was the offense?

In [73]:
```
df2 = pd.read_csv('PDI_Police_Data_Initiative_Crime_Incidents.csv')
```

In [79]:
```
df2['DATE_REPORTED'].min(), df2['DATE_REPORTED'].max()
```

Out[79]:
```
('01/01/2022 01:08:00 AM', '06/26/2022 12:50:00 AM')
```

In [81]:
```
df2['SUSPECT_AGE'].value_counts()
```

Out[81]:
```
SUSPECT_AGE
UNKNOWN     9003
18-25       1778
31-40       1525
26-30       1126
41-50        659
UNDER 18     629
51-60        298
61-70        121
OVER 70       16
Name: count, dtype: int64
```

In [137…
```
#b
df2.shape
```

Out[137…
```
(15155, 40)
```

In [143…
```
#c 'OPENING' Column has most missing values
df2.isna().sum()
```

Out[143…
```
INSTANCEID                         0
INCIDENT_NO                        0
DATE_REPORTED                      0
DATE_FROM                          2
DATE_TO                            9
CLSD                             545
UCR                               10
DST                                0
BEAT                              28
OFFENSE                           10
LOCATION                           2
THEFT_CODE                     10167
FLOOR                          14127
SIDE                           14120
OPENING                        14508
HATE_BIAS                          0
DAYOFWEEK                        423
RPT_AREA                         239
CPD_NEIGHBORHOOD                 249
WEAPONS                            5
DATE_OF_CLEARANCE               2613
HOUR_FROM                          2
HOUR_TO                            9
ADDRESS_X                        148
LONGITUDE_X                     1714
LATITUDE_X                      1714
VICTIM_AGE                         0
VICTIM_RACE                     2192
VICTIM_ETHNICITY                2192
VICTIM_GENDER                   2192
SUSPECT_AGE                        0
SUSPECT_RACE                    7082
SUSPECT_ETHNICITY               7082
SUSPECT_GENDER                  7082
TOTALNUMBERVICTIMS                33
TOTALSUSPECTS                   7082
UCR_GROUP                         10
ZIP                                1
COMMUNITY_COUNCIL_NEIGHBORHOOD  1639
SNA_NEIGHBORHOOD                1632
dtype: int64
```

In [151…
```
#d '04/03/2022' has the highest frequency
df2['DATE_REPORTED'].describe()
```

Out[151…
```
count                      15155
unique                     10961
top       04/03/2022 12:59:00 PM
freq                          36
Name: DATE_REPORTED, dtype: object
```

In [155…
```
#e DATE: 12/31/2001-1/02/2022?? OFFENSE: THEFT FROM MOTOR VEHICLE
df2.loc[100]
```

Out[155…
```
INSTANCEID                        33153491-2F02-4643-8342-F07153D413F6
INCIDENT_NO                                                  229000105
DATE_REPORTED                                   01/02/2022 11:25:00 AM
DATE_FROM                                       12/31/2001 02:00:00 PM
DATE_TO                                         01/02/2022 11:20:00 AM
CLSD                                                   Z--EARLY CLOSED
UCR                                                              600.0
DST                                                                  5
BEAT                                                                 4
OFFENSE                                                          THEFT
LOCATION                                    02-MULTI FAMILY APARTMENT
THEFT_CODE                                 23F-THEFT FROM MOTOR VEHICLE
FLOOR                                                              NaN
SIDE                                                               NaN
OPENING                                                            NaN
HATE_BIAS                                  N--NO BIAS/NOT APPLICABLE
DAYOFWEEK                                                       MONDAY
RPT_AREA                                                           432
CPD_NEIGHBORHOOD                                           MOUNT AIRY
WEAPONS                                                       99 - NONE
DATE_OF_CLEARANCE                               02/10/2022 12:00:00 AM
HOUR_FROM                                                        140.0
HOUR_TO                                                         1120.0
ADDRESS_X                                            25XX FLANIGAN CT
LONGITUDE_X                                                 -84.574648
LATITUDE_X                                                   39.203036
VICTIM_AGE                                                       31-40
VICTIM_RACE                                                      BLACK
VICTIM_ETHNICITY                                   NOT OF HISPANIC ORIG
VICTIM_GENDER                                                   FEMALE
SUSPECT_AGE                                                      31-40
SUSPECT_RACE                                                     BLACK
SUSPECT_ETHNICITY                                  NOT OF HISPANIC ORIG
SUSPECT_GENDER                                                    MALE
TOTALNUMBERVICTIMS                                                 1.0
TOTALSUSPECTS                                                      1.0
UCR_GROUP                                                        THEFT
ZIP                                                            45239.0
COMMUNITY_COUNCIL_NEIGHBORHOOD                              MOUNT AIRY
SNA_NEIGHBORHOOD                                               MT. AIRY
Name: 100, dtype: object
```

In [ ]: