Applied exercise

March 7, 2022

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[3]:
```
df= pd.read_csv("../Downloads/claimsfile.csv")
```

```
C:\Users\narae\anaconda3\lib\site-
packages\IPython\core\interactiveshell.py:3165: DtypeWarning: Columns (0,11,12)
have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

[4]:
```
df.head()
```

[4]:

	Claim Number	Date Received	Incident Date	Airport Code	Airport Name	Airline Name	Claim Type	Claim Site	Item	Claim Amount	Status	Close Amount	Disposition
0	0909802M	4-Jan-02	12/12/2002 0:00	EWR	Newark International Airport	Continental Airlines	Property Damage	Checkpoint	Other	$350.00	Approved	$350.00	Approve in Full
1	0202417M	2-Feb-02	1/16/2004 0:00	SEA	Seattle-Tacoma International	NaN	Property Damage	Checked Baggage	Luggage (all types including footlockers)	$100.00	Settled	$50.00	Settle
2	0202445M	4-Feb-02	11/26/2003 0:00	STL	Lambert St. Louis International	American Airlines	Property Damage	Checked Baggage	Cell Phones	$278.88	Settled	$227.92	Settle
3	0909816M	7-Feb-02	1/6/2003 0:00	MIA	Miami International Airport	American Airlines	Property Damage	Checkpoint	Luggage (all types including footlockers)	$50.00			
4	2005032379513	18-Feb-02	2/5/2005 0:00	MCO	Orlando International Airport	Delta (Song)	Property Damage	Checkpoint	Baby - Strollers; car seats; playpen; etc.	$84.79			

```
3  Approved      $50.00   Approve in Full
4  Approved      $84.79   Approve in Full
```

[6]:
```
df.drop_duplicates(inplace=True)
```

[7]:
```
df.isnull().sum()
```

[7]:
```
Claim Number         0
Date Received      263
Incident Date     2183
Airport Code      8523
Airport Name      8523
Airline Name     34373
Claim Type        7913
Claim Site         740
Item              3966
Claim Amount      4043
Status               5
Close Amount     68951
Disposition      72907
dtype: int64
```

[8]:
```
df.dropna(axis=0,thresh=2,inplace=True)
```

[9]:
```
df=df.replace('^-',np.nan,regex=True)
```

[10]:
```
df.isnull().sum().sum()
```

[10]:
```
247944
```

[11]:
```
df.columns
```

[11]:
```
Index(['Claim Number', 'Date Received', 'Incident Date', 'Airport Code',
       'Airport Name', 'Airline Name', 'Claim Type', 'Claim Site', 'Item',
       'Claim Amount', 'Status', 'Close Amount', 'Disposition'],
      dtype='object')
```

[12]:
```
df['Date Received'].fillna(method='ffill', inplace=True)
df['Incident Date'].fillna(method='ffill', inplace=True)
df['Incident Date'].fillna(method='ffill', inplace=True)
df['Airport Code'].fillna('Unknown value',inplace=True)
df['Airport Name'].fillna("Unknown",inplace=True)
df['Airline Name'].fillna("Unknown",inplace=True)
df['Claim Type'].fillna("Unknown",inplace=True)
df['Claim Site'].fillna("Unknown",inplace=True)
df['Item'].fillna("Unknown",inplace=True)
df['Claim Amount'].fillna(0.00,inplace=True)
```

```
df['Close Amount'].fillna(0.00,inplace=True)
df['Disposition'].fillna("Cancelled",inplace=True)
df["Status"].fillna("Unknown",inplace=True)
```

[13]:
```
df.head()
```

[13]:
```
     Claim Number Date Received    Incident Date Airport Code  \
0       0909802M      4-Jan-02  12/12/2002 0:00          EWR
1       0202417M      2-Feb-02   1/16/2004 0:00          SEA
2       0202445M      4-Feb-02  11/26/2003 0:00          STL
3       0909816M      7-Feb-02    1/6/2003 0:00          MIA
4  2005032379513     18-Feb-02    2/5/2005 0:00          MCO

                       Airport Name          Airline Name       Claim Type  \
0     Newark International Airport  Continental Airlines  Property Damage
1     Seattle-Tacoma International               Unknown  Property Damage
2  Lambert St. Louis International     American Airlines  Property Damage
3      Miami International Airport     American Airlines  Property Damage
4    Orlando International Airport           Delta (Song)  Property Damage

        Claim Site                                        Item Claim Amount  \
0       Checkpoint                                       Other     $350.00
1  Checked Baggage   Luggage (all types including footlockers)     $100.00
2  Checked Baggage                                 Cell Phones     $278.88
3       Checkpoint   Luggage (all types including footlockers)      $50.00
4       Checkpoint  Baby - Strollers; car seats; playpen; etc.      $84.79

     Status Close Amount      Disposition
0  Approved     $350.00   Approve in Full
1   Settled      $50.00            Settle
2   Settled     $227.92            Settle
3  Approved      $50.00   Approve in Full
4  Approved      $84.79   Approve in Full
```

[14]:
```
df.tail()
```

[14]:
```
          Claim Number Date Received Incident Date   Airport Code Airport Name  \
204262  2015120427297     20-Nov-15     16-Oct-15  Unknown value      Unknown
204263  2015123027969     17-Dec-15      2-Dec-15  Unknown value      Unknown
204264  2016010428072     22-Dec-15     20-Dec-15  Unknown value      Unknown
204265  2016011328300     30-Dec-15     28-Dec-15  Unknown value      Unknown
204266  2015123128015     31-Dec-15     23-Nov-15  Unknown value      Unknown

       Airline Name                Claim Type       Claim Site  \
204262      Unknown           Property Damage  Checked Baggage
204263      Unknown           Property Damage  Checked Baggage
204264      Unknown  Passenger Property Loss  Checked Baggage
```

```
204265      Unknown  Passenger Property Loss  Checked Baggage
204266      Unknown  Passenger Property Loss       Checkpoint

                                                    Item Claim Amount  \
204262  Baggage/Cases/Purses; Books; Magazines & Other…          0.0
204263                            Audio/Video; Home Decor          0.0
204264                                           Clothing          0.0
204265                 Tools & Home Improvement Supplies          0.0
204266                                Personal Accessories          0.0

         Status Close Amount Disposition
204262  Unknown          0.0   Cancelled
204263  Unknown          0.0   Cancelled
204264  Unknown          0.0   Cancelled
204265  Unknown          0.0   Cancelled
204266  Unknown          0.0   Cancelled
```

[15]:
```
# Handling special characters in between
#train_num = train_num.fillna(train_num.median(
df=df.replace('\$|;|,','',regex=True).astype(str)
df['Claim Amount']=df['Claim Amount'].replace('nan','0.00',regex=True).
 ↪astype(str)
#df=df.astype({'Claim Amount':np.float, 'Close Amount':np.float})
df.head()
```

[15]:
```
    Claim Number Date Received    Incident Date Airport Code  \
0       0909802M      4-Jan-02  12/12/2002 0:00          EWR
1       0202417M      2-Feb-02   1/16/2004 0:00          SEA
2       0202445M      4-Feb-02  11/26/2003 0:00          STL
3       0909816M      7-Feb-02    1/6/2003 0:00          MIA
4  2005032379513     18-Feb-02    2/5/2005 0:00          MCO

                      Airport Name          Airline Name       Claim Type  \
0     Newark International Airport  Continental Airlines  Property Damage
1     Seattle-Tacoma International               Unknown  Property Damage
2  Lambert St. Louis International     American Airlines  Property Damage
3      Miami International Airport     American Airlines  Property Damage
4    Orlando International Airport           Delta (Song)  Property Damage

        Claim Site                                      Item Claim Amount  \
0       Checkpoint                                     Other      350.00
1  Checked Baggage  Luggage (all types including footlockers)      100.00
2  Checked Baggage                                Cell Phones      278.88
3       Checkpoint  Luggage (all types including footlockers)       50.00
4       Checkpoint    Baby - Strollers car seats playpen etc.       84.79

     Status Close Amount      Disposition
```

```
0  Approved      350.00   Approve in Full
1   Settled       50.00            Settle
2   Settled      227.92            Settle
3  Approved       50.00   Approve in Full
4  Approved       84.79   Approve in Full
```

[31]:
```
# Converting amounts to float and date values to datetime

df=df.astype({'Claim Amount':float, 'Close Amount':float})
df['Incident Date']=pd.to_datetime(df['Incident Date'],errors = 'coerce')
df['Date Received']=pd.to_datetime(df['Date Received'],errors = 'coerce')
```

[32]:
```
type(df['Incident Date'].values[0])
```

[32]:
```
numpy.datetime64
```

[33]:
```
type(df['Date Received'].values[0])
```

[33]:
```
numpy.datetime64
```

[17]:
```
df.describe()
```

[17]:
```
       Claim Amount  Close Amount
count  2.042610e+05  204261.00000
mean   1.477336e+07      65.11558
std    6.637890e+09     754.98190
min    0.000000e+00       0.00000
25%    1.095000e+01       0.00000
50%    9.842000e+01       0.00000
75%    3.090000e+02      30.00000
max    3.000000e+12  250000.00000
```

[18]:
```
closeamt_array=df['Close Amount'].to_numpy()
closeamt_array
```

[18]:
```
array([350.  ,  50.  , 227.92, …,   0.  ,   0.  ,   0.  ])
```

[19]:
```
##1. What is the most common type of insurance claim?
import statistics
from statistics import mode
a=df['Claim Type'].to_list()
print(mode(a))
```

```
Passenger Property Loss
```

[20]:
```
##2. Which claim site within the airport are claims most commonly filed for?
a=df['Claim Site'].to_list()
```

```
print(mode(a))
```

```
Checked Baggage
```

[21]:
```
##3. What type of claim is made most at each claim site?
a=df['Claim Type']
a.index=df['Claim Site']
print(mode(a['Checkpoint']))
print(mode(a['Checked Baggage']))
print(mode(a['Motor Vehicle']))
print(mode(a['Bus Station']))
```

```
Property Damage
Passenger Property Loss
Motor Vehicle
Passenger Property Loss
```

[22]:
```
##4. What is the typical claim amount?
from collections import Counter
counted_list= Counter(list(df['Claim Amount']))
top_claim_amts=counted_list.most_common(3)
top_claim_amts
```

[22]:
```
[(0.0, 45235), (100.0, 3871), (50.0, 3460)]
```

[23]:
```
##5. What is the overall claim approval rate for the entire U.S.?

a=df['Status'].to_list()
a.count('Approved')/len(a)
```

[23]:
```
0.17134450531427928
```

[24]:
```
##6. If a claim is approved or settled, what percent of the claim amount do the␣
 ↪airports pay?
new_list = list(zip(df.loc[:]['Close Amount'], df.loc[:]["Status"],df.loc[:
 ↪]["Claim Amount"]))
```

Here, we take the total of the claim amount when the status is approved or settled. To get the amount that the airport pays, we subtract the close amount from the claim amount as the airport pays the difference. Then we take the percentage of it.

[27]:
```
total_claim=0
paid_by_airport=0
for cls_amt,status,clm in new_list:
    if status=="Approved" or status=="Settled":
        total_claim+=clm
        paid_by_airport+=clm-cls_amt
```

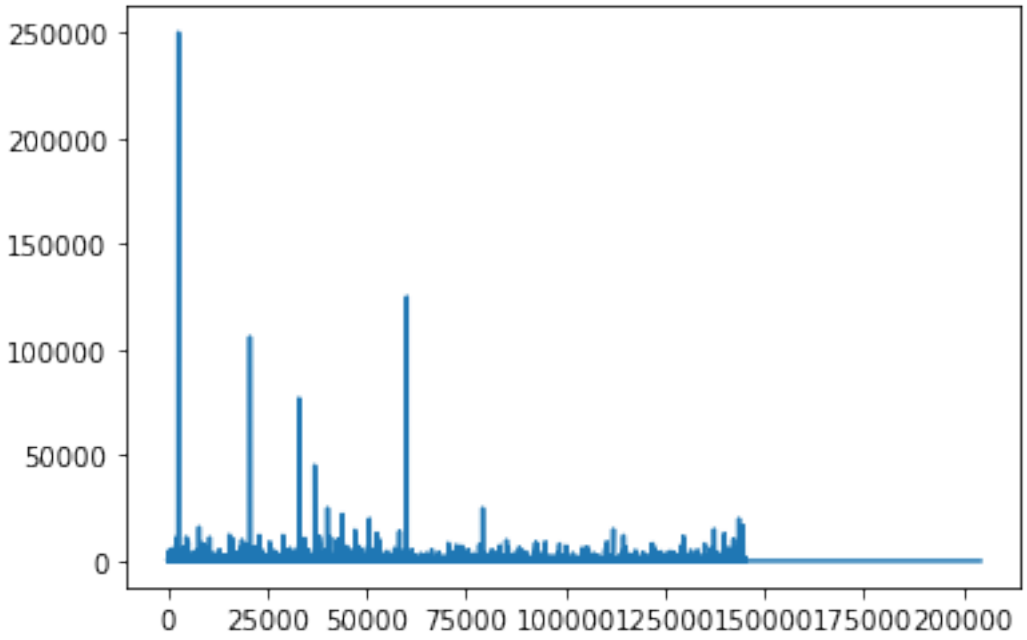
[28]:
```
(paid_by_airport/total_claim)*100
```

[28]:
```
55.89950510373774
```

[29]:
```
##7. What are the five airports with the most claims?
df['Airport Name'].value_counts().head(6)
```

[29]:
```
John F. Kennedy International           9232
Unknown                                 8959
Los Angeles International Airport       7260
Newark International Airport            6866
Chicago O'Hare International Airport    6843
Miami International Airport             6432
Name: Airport Name, dtype: int64
```

[30]:
```
##8. Has the total close amount increased or decreased over time?
plt.plot(np.array(df["Close Amount"].index), closeamt_array)
plt.show()
```



When we look at the graph, we can see that the **close amount decreases over time**. How can we say that when we did not plot it using the dates or using time series analysis? When we look at the data, we can see that the claim dates are arranged in order from the year 2002 to 2015.

[ ]:

[ ]: