Assignment 1 FIT1043

Name: Ekramul Islam

ID: 33275548  
Submission Date: Monday, 29th August, 2022

**Task A: Data Exploration And Auditing:**

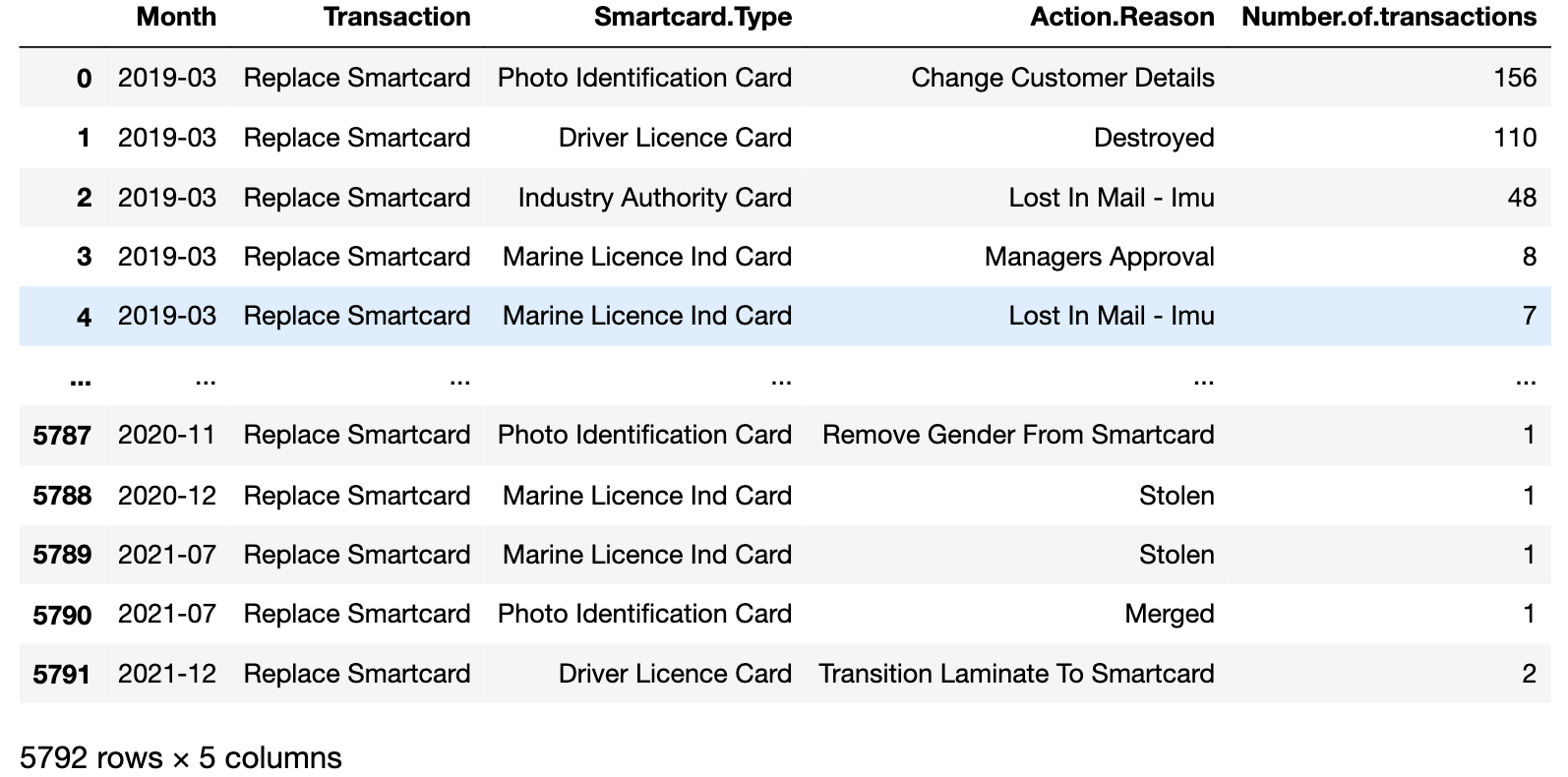
**A1: Dataset Size**

**Codes:**

smartcard = pd.read\_csv('monthly\_smartcard\_replacements.csv')

smartcard

**Output:**



We can see from the table above that there exists 5792 rows and 5 columns of data instances and variables

**A2: Missing values in the dataset**

**Codes:**

smartcard.isnull().sum()

**Text

Description automatically generatedOutput**:

As seen from the above table, there are no missing datasets.

**A3: Data types**

**Codes:**

smartcard.dtypes

Output:

Graphical user interface, text, application

Description automatically generated

As seen from the above table, Month, Transaction, Smartcard.Type and Action.Reason have the **datatype object**. Number.of.transactions has the **datatype int64.**

**A4: Convert Data Type**

**Codes:**

smartcard ['Month'] = pd.to\_datetime(smartcard['Month'])

smartcard

Table

Description automatically generated

**Output:**

**A5: Descriptive Statistics**

**Codes:**

smartcard.describe()

**Output:**

Table

Description automatically generated

**Some Observations:**

* The **standard deviation** for the number.of.transactions is high. This means that the data points are dispersed and not mostly centered.
* We can see fromThe **count** tells us that there are in total of 5792 number of data points in the number.of.transactions column.
* We can draw conclusion from **max** that the maximum number.of.transactions was 9097.
* We see that the **25%** of number.of.transactions was around 5 times, **50%** was around 18 times, and **75%** was around 84 times.

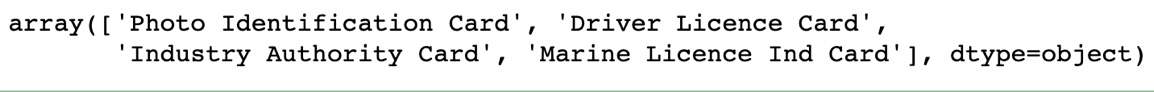
**A6: Exploring Smartcard Types**

1 )

**Codes:**

smartcard['Smartcard.Type'].unique()

**Output:**



There are four different unique smartcard types recorded in the ‘Smartcard.Type’ column. Those different smartcard types are as follows: 'Photo Identification Card' 'Driver Licence Card','Industry Authority Card' 'Marine Licence Ind Card'.

2 )

**Codes:**

u = smartcard['Smartcard.Type']

c = u.value\_counts()

c

p = u.value\_counts(normalize=True)

p

pf=u.value\_counts(normalize=True).mul(100).round(1).astype(str)+'%'

pf

**Output:**

Text

Description automatically generated

The percentage of Driver Licence Card records as one of the smartcard types in ‘Smartcard.Type’ column is 32.7%.

**A7: Exploring reasons for Smartcard Replacement**

1 )

**Code:**

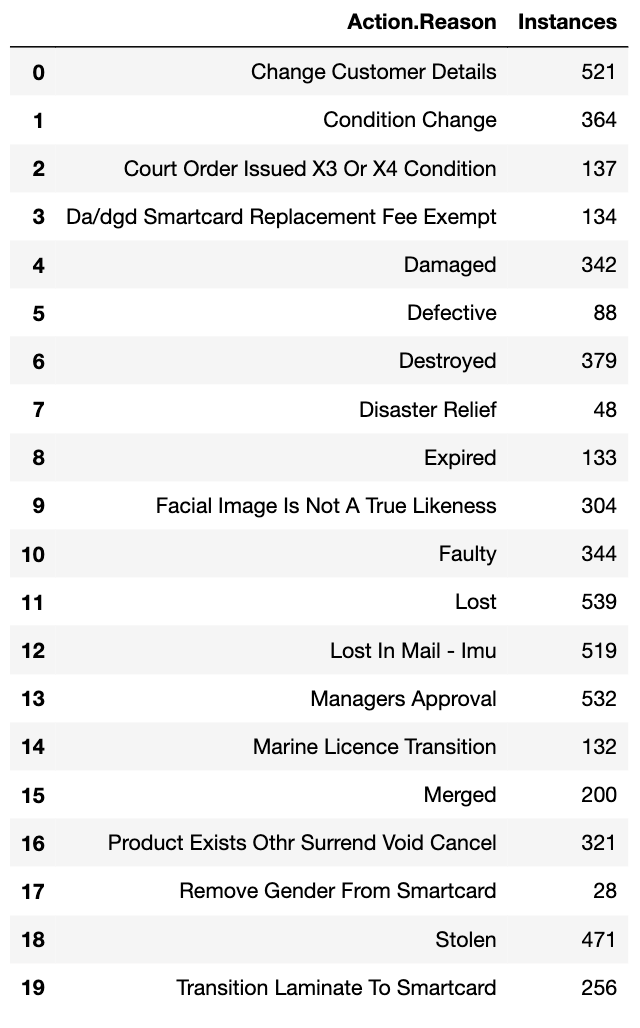
different\_reasons = smartcard.groupby(["Action.Reason"]).count().reset\_index()

df\_new = different\_reasons[['Action.Reason', 'Month']]

df\_new.rename(columns = {'Month':'Instances'}, inplace = True)

df\_new

**Output:**



There are in total 20 reasons for smartcard replacement. The different reasons are given under the **Action.Reason** and the number of instances observed are given under the **Instances** column in the Output above.

2 )

**Code:**

filt = (smartcard['Action.Reason'] == 'Lost') & (smartcard['Number.of.transactions'] >= 100)

smartcard[filt]

**Output:**

Table

Description automatically generated

The total number of month in which 100 or more smartcard replacement are reported being lost is 264.

**Task B: Group level analysis and visualisation:**

**B1: Investigating Annual Smartcard Replacements**

1 )

**Code:**

smartcard['Year']=smartcard['Month'].dt.year

smartcard

**Output:**

Table

Description automatically generated

2 )

**Code:**

Trans\_year=smartcard.drop(['Month','Transaction','Smartcard.Type','Action.Reason'],axis=1)

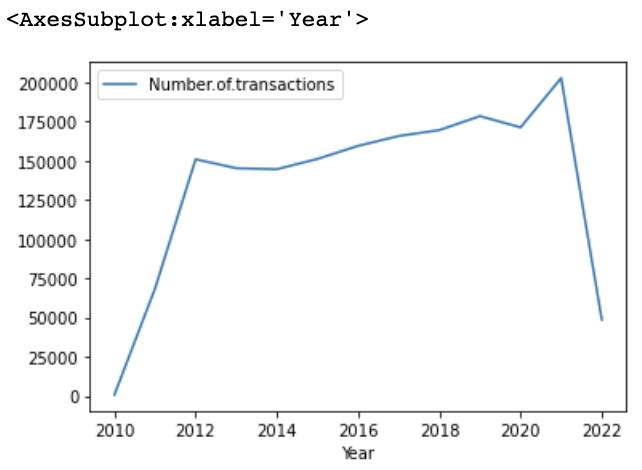
Trans\_year

TY=Trans\_year.groupby('Year').sum()

TY

TY.plot()

**Output:**



3 )

W

3 )

From the chart, we see that the number of transactions increased rapidly from 2010 to 2012. After that, it stayed fairly same by fluctuating up and down by a little margin. From 2020 to 2021, there is a final steep incline of number of transaction. Finally, from 2021 to 2022, we see a very steep decline in the number of transaction.

**B2: Investigating Reasons For Smartcard Replacements**

1 )

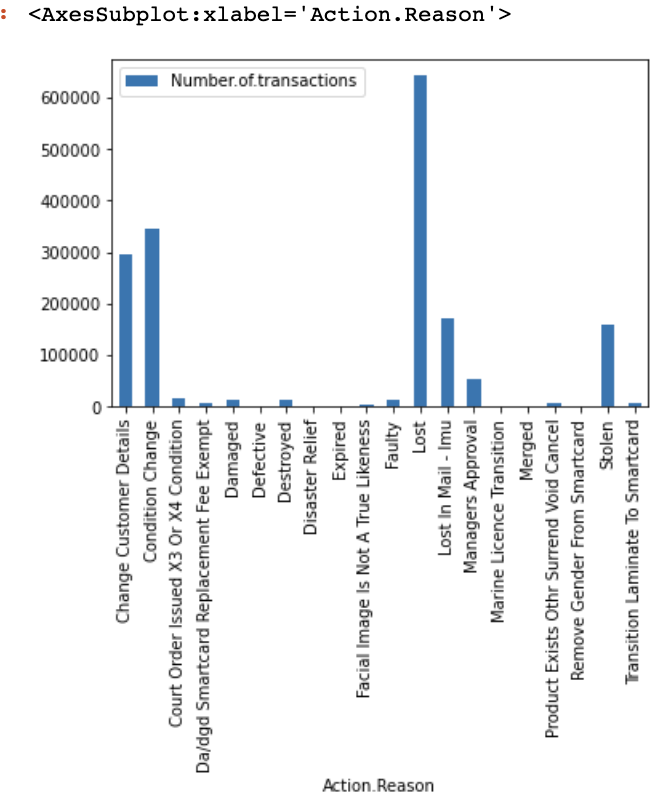
**Code:**

df\_r\_bar=smartcard.drop(['Month','Transaction','Smartcard.Type','Year'], axis=1)

df\_r\_bar

df5=df\_r\_bar.groupby('Action.Reason').sum()

df5.plot(kind='bar',rot=90)

**Output:**

2 )

**Code:**

df5['Number.of.transactions'].nlargest(n=5)

Text, letter

Description automatically generated**Output:**

As we see from the data above, the top three smartcard replacement reasons are Lost, Condition Change, and Change Customer Details.

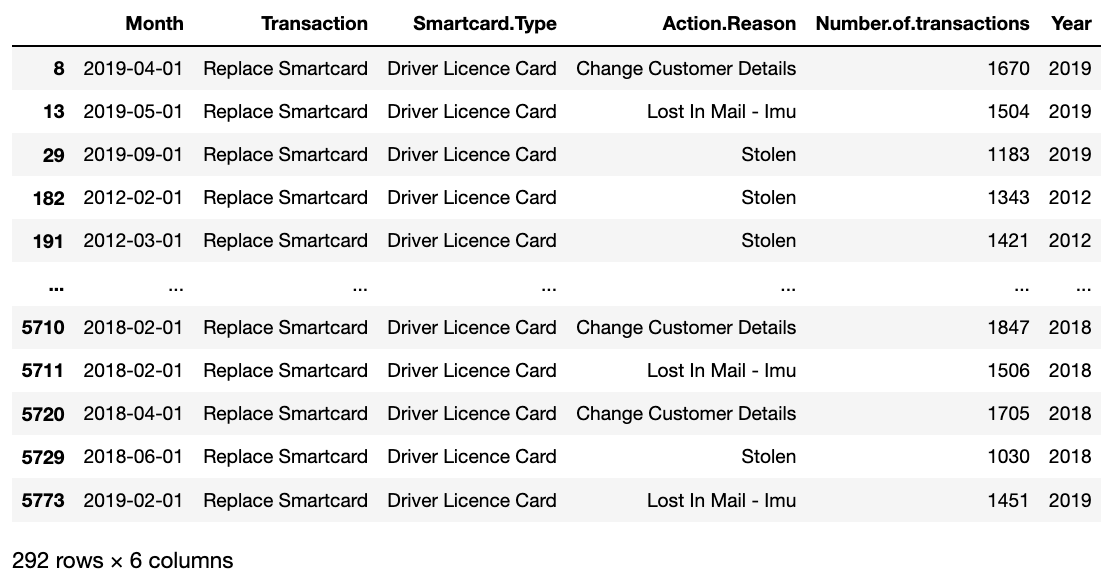
3 )

**Code:**

df6=smartcard.loc[(smartcard['Number.of.transactions']<2000)&(smartcard['Number.of.transactions']>1000)]

df6

**Output:**



We can see from the above table that there are 292 total number of transactions of which ‘Action.Reason’ is between 1000 and 2000.

**B3. Investigating reasons over annual smartcard replacement**

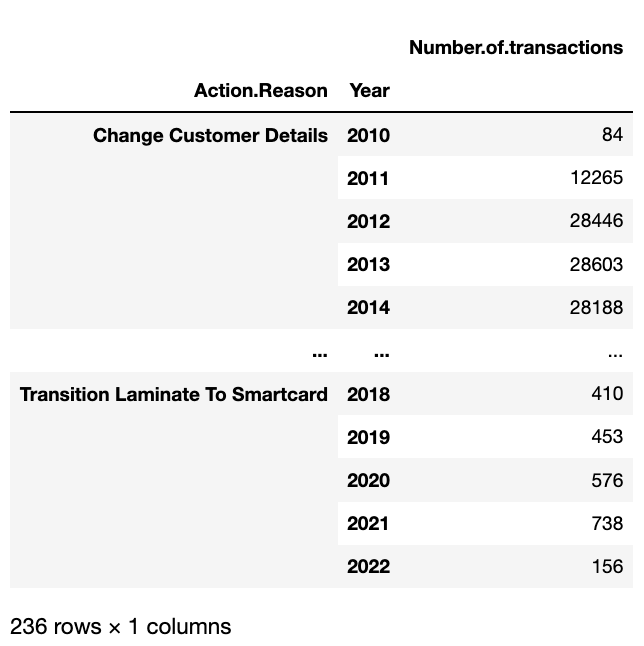
1 )

**Codes:**

df7={'Number.of.transactions':'sum'}

df8=smartcard.groupby(['Action.Reason','Year']).agg(df7)

df8

**Output**:

2 )

**Code:**

filt=df8[df8['Number.of.transactions']>10000]

filt

filt.count()

**Output:**

A picture containing text

Description automatically generated

Therefore, the number of years is 52.

3 )

**Code:**

filt=df8[df8['Number.of.transactions']>10000]

filt

**Output:**

The action reasons that have at least one

Year with number of annual transactions

Exceeding 10000 are :

1. Change customer details
2. Condition change
3. Lost
4. Lost in mail – Imu
5. Stolen

4 )

**Code:**

df8.hist()

**Output:**

Table

Description automatically generated

5 )

Most of the number of transactions per action reason is below 10000.

The histogram is Right Skewed.