

2000080110_ML SKILL1

September 2, 2021

```
[15]: import pandas as pd
import numpy as np
df=pd.read_csv(r'E:\M&L excel\M&L Skill1.csv')
df
#rename
df=df.rename(columns={"A":"TransactionID", "B":"Transaction Date", "C":
    ↳ "UserID", "D":"ProductID", "E":"Quantity"})
df
```

```
[15]:
```

	TransactionID	Transaction Date	UserID	ProductID	Quantity
0	1	21-08-2010	7.0	2	1
1	2	26-05-2011	NaN	4	1
2	3	16-06-2011	3.0	3	1
3	4	26-08-2012	1.0	2	3
4	5	06-06-2013	NaN	4	1
5	6	23-12-2013	2.0	5	6
6	7	30-12-2013	3.0	4	1
7	8	24-04-2014	NaN	2	3
8	9	24-04-2015	7.0	4	3
9	10	08-05-2016	3.0	4	4

```
[16]: import pandas as pd
import numpy as np
#Order the rows of transactions by Transaction_ID in descending order
df.sort_values(by=['TransactionID'],ascending=[False])
#in order to totally change data into sorted format we need to re-assign it to
↳ 'df'
```

```
[16]:
```

	TransactionID	Transaction Date	UserID	ProductID	Quantity
9	10	08-05-2016	3.0	4	4
8	9	24-04-2015	7.0	4	3
7	8	24-04-2014	NaN	2	3
6	7	30-12-2013	3.0	4	1
5	6	23-12-2013	2.0	5	6
4	5	06-06-2013	NaN	4	1
3	4	26-08-2012	1.0	2	3
2	3	16-06-2011	3.0	3	1
1	2	26-05-2011	NaN	4	1

0	1	21-08-2010	7.0	2	1
---	---	------------	-----	---	---

```
[17]: #Order the rows of transactions by Quantity ascending, TransactionDate_
      ↪descending
      df.sort_values(by=['Quantity', 'Transaction Date'], ascending=[True, False])
```

```
[17]:
```

	TransactionID	Transaction Date	UserID	ProductID	Quantity
6	7	30-12-2013	3.0	4	1
1	2	26-05-2011	NaN	4	1
0	1	21-08-2010	7.0	2	1
2	3	16-06-2011	3.0	3	1
4	5	06-06-2013	NaN	4	1
3	4	26-08-2012	1.0	2	3
8	9	24-04-2015	7.0	4	3
7	8	24-04-2014	NaN	2	3
9	10	08-05-2016	3.0	4	4
5	6	23-12-2013	2.0	5	6

```
[18]: #Display only the row number 2, 4, 7
      print("DATA---")
      print(df)
      print(df.iloc[2])
      print(df.iloc[4])
      print(df.iloc[7])
      #Count the total missing values in the data set.
      print("\nTotal missing values are--", df.isna().sum().sum())
```

DATA---

	TransactionID	Transaction Date	UserID	ProductID	Quantity
0	1	21-08-2010	7.0	2	1
1	2	26-05-2011	NaN	4	1
2	3	16-06-2011	3.0	3	1
3	4	26-08-2012	1.0	2	3
4	5	06-06-2013	NaN	4	1
5	6	23-12-2013	2.0	5	6
6	7	30-12-2013	3.0	4	1
7	8	24-04-2014	NaN	2	3
8	9	24-04-2015	7.0	4	3
9	10	08-05-2016	3.0	4	4

TransactionID 3

Transaction Date 16-06-2011

UserID 3.0

ProductID 3

Quantity 1

Name: 2, dtype: object

TransactionID 5

Transaction Date 06-06-2013

UserID NaN

```

ProductID          4
Quantity           1
Name: 4, dtype: object
TransactionID       8
Transaction Date    24-04-2014
UserID             NaN
ProductID          2
Quantity           3
Name: 7, dtype: object

```

Total missing values are-- 3

```

[19]: #Replace the missing values with the average of the column where missing values
      ↪are present
df=df.fillna(df.mean())
print(df)
print()
#Print all the transaction where quantity is greater than 1
print(df[df.iloc[:,4]>1])

```

	TransactionID	Transaction Date	UserID	ProductID	Quantity
0	1	21-08-2010	7.000000	2	1
1	2	26-05-2011	3.714286	4	1
2	3	16-06-2011	3.000000	3	1
3	4	26-08-2012	1.000000	2	3
4	5	06-06-2013	3.714286	4	1
5	6	23-12-2013	2.000000	5	6
6	7	30-12-2013	3.000000	4	1
7	8	24-04-2014	3.714286	2	3
8	9	24-04-2015	7.000000	4	3
9	10	08-05-2016	3.000000	4	4

	TransactionID	Transaction Date	UserID	ProductID	Quantity
3	4	26-08-2012	1.000000	2	3
5	6	23-12-2013	2.000000	5	6
7	8	24-04-2014	3.714286	2	3
8	9	24-04-2015	7.000000	4	3
9	10	08-05-2016	3.000000	4	4

```

[20]: #Remove the column Transaction_ID and print the dataframe
df.drop(['TransactionID'],axis=1)

```

```

[20]: Transaction Date    UserID  ProductID  Quantity
0      21-08-2010    7.000000         2         1
1      26-05-2011    3.714286         4         1
2      16-06-2011    3.000000         3         1
3      26-08-2012    1.000000         2         3
4      06-06-2013    3.714286         4         1

```

5	23-12-2013	2.000000	5	6
6	30-12-2013	3.000000	4	1
7	24-04-2014	3.714286	2	3
8	24-04-2015	7.000000	4	3
9	08-05-2016	3.000000	4	4

```
[21]: #Read the dataset in it's initial from (having missing values). Drop the rows
      ↪having missing values and print the data set
df1=pd.read_csv('E:\M&L excel\M&L Skill1.csv')
df1.dropna()#reassign if want to remove permanently
```

```
[21]:
```

	A	B	C	D	E
0	1	21-08-2010	7.0	2	1
2	3	16-06-2011	3.0	3	1
3	4	26-08-2012	1.0	2	3
5	6	23-12-2013	2.0	5	6
6	7	30-12-2013	3.0	4	1
8	9	24-04-2015	7.0	4	3
9	10	08-05-2016	3.0	4	4

```
[22]: #Print what is the total quality in the dataset
l=list(df)
l.remove('TransactionID')
l.remove('Transaction Date')
l.remove('UserID')
l.remove('ProductID')
l
x=df[l].sum(axis=1).sum()
print("Total quantity is",x)
```

Total quantity is 24