13 nth October Notes

Importing and Reading CSV Files

import pandas as pd

sale1=pd.read_csv("/home/Sales Transactions-2017.csv")

sale1

```
#13/10/25
import pandas as pd
sale1=pd.read_csv("/home/Sales Transactions-2017.csv")
sale1
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
0	1/4/2017	Sal:1	SOLANKI PLASTICS	DONA-VAI-9100	2	1,690.00	3,380.00	NaN	13,100.00
1	1/4/2017	Sal:1	SOLANKI PLASTICS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
2	1/4/2017	Sal:2	SARNESWARA TRADERS	VISHNU CHOTA WINE	500	23	11,500.00	NaN	30,990.00
3	1/4/2017	Sal:2	SARNESWARA TRADERS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
4	1/4/2017	Sal:2	SARNESWARA TRADERS	DONA-VAI-9100	5	1,690.00	8,450.00	NaN	NaN
...	...	...	...	...	...	...	...	...	...
47285	31/03/2018	Sal:10042	Vkp	10*10 SHEET	25	137	3,425.00	NaN	3,425.00
47286	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
47287	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
47288	NaN	Total	NaN	NaN	607,734.60	669,300.49	9,953,816.13	106,607.00	9,868,583.13
47289	NaN	Total	NaN	NaN	7,593,062.00	8,309,116.00	115,778,725.71	936,348.00	115,105,123.71

47290 rows × 9 columns

Combining Tables

✅ Rules for Combining Tables

1. All tables must have the same number of columns
2. The order of columns should be the same in all tables

🔹 Combining Tables

sale2=pd.read_csv("/home/Sales Transactions-2018.csv")

sale2

```
sale2=pd.read_csv("/home/Sales Transactions-2018.csv")
sale2
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
0	1/4/2018	Sal:146	TP13	SILVER POUCH 9*12	50	85	4,250.00	NaN	66,724.00
1	1/4/2018	Sal:146	TP13	RUBBER	5	290	1,450.00	NaN	NaN
2	1/4/2018	Sal:146	TP13	DURGA 10*12 Blue	1,600.00	5.5	8,800.00	NaN	NaN
3	1/4/2018	Sal:146	TP13	DURGA 13*16 BLUE	400	11	4,400.00	NaN	NaN
4	1/4/2018	Sal:146	TP13	10*12 SARAS-NAT	600	8.1	4,860.00	NaN	NaN
...	...	...	...	...	...	...	...	...	...
44735	31/03/2019	Sal:9610	HAMPI FOODS	SPOON SOOFY	200	40	8,000.00	NaN	NaN
44736	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
44737	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
44738	NaN	Total	NaN	NaN	666,056.00	1,067,808.80	10,796,991.30	29,999.00	10,787,647.30
44739	NaN	Total	NaN	NaN	7,097,803.00	10,024,197.00	117,897,671.80	720,204.00	117,427,983.80

44740 rows × 9 columns

sale3=pd.read_csv("/home/Sales Transactions-2019.csv")

sale3

```
sale3=pd.read_csv("/home/Sales Transactions-2019.csv")
sale3
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
0	1/4/2019	Sal:687	BALAJI PLASTICS	DONA-VAI-9100	1	1,730.00	1,730.00	NaN	3,460.00
1	1/4/2019	Sal:687	BALAJI PLASTICS	SMART BOUL(48)	1	1,730.00	1,730.00	NaN	NaN
2	1/4/2019	Sal:688	BALAJI PLASTICS	Vishnu Ice	110	18.5	2,035.00	NaN	2,035.00
3			28/3		0	0			
4	1/4/2019	Sal:689	BALAJI PLASTICS	100LEAF -SP	3	585	1,755.00	NaN	1,755.00
...	...	...	...	...	...	...	...	...	...
19171	10/10/2019	Sal:4935	K.SRIHARI	13*16 WHITE RK	400	16	6,400.00	NaN	NaN
19172	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19173	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19174	NaN	Total	NaN	NaN	99,284.90	175,381.65	2,203,649.50	20,680.00	2,189,014.50
19175	NaN	Total	NaN	NaN	2,710,193.00	5,519,888.40	53,360,791.40	672,984.00	52,830,224.40

19176 rows × 9 columns

#combining table rules same number of coloumn in all 3 table,order of table should be same

#if all the coloumn has same number of coloumn we combine that tables

final_data=pd.concat([sale1,sale2,sale3],ignore_index=True)

final_data

```
#combining table rules same number of coloumn in all 3 table,order of table should be same
#if all the coloumn has same number of coloumn we combine that tables
final_data=pd.concat([sale1,sale2,sale3],ignore_index=True)
final_data
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
0	1/4/2017	Sal:1	SOLANKI PLASTICS	DONA-VAI-9100	2	1,690.00	3,380.00	NaN	13,100.00
1	1/4/2017	Sal:1	SOLANKI PLASTICS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
2	1/4/2017	Sal:2	SARNESWARA TRADERS	VISHNU CHOTA WINE	500	23	11,500.00	NaN	30,990.00
3	1/4/2017	Sal:2	SARNESWARA TRADERS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
4	1/4/2017	Sal:2	SARNESWARA TRADERS	DONA-VAI-9100	5	1,690.00	8,450.00	NaN	NaN
...	...	...	...	...	...	...	...	...	...
111201	10/10/2019	Sal:4935	K.SRIHARI	13*16 WHITE RK	400	16	6,400.00	NaN	NaN
111202	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
111203	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
111204	NaN	Total	NaN	NaN	99,284.90	175,381.65	2,203,649.50	20,680.00	2,189,014.50
111205	NaN	Total	NaN	NaN	2,710,193.00	5,519,888.40	53,360,791.40	672,984.00	52,830,224.40

111206 rows × 9 columns

Initial Data Analysis

Shape of the Combined Data

final_data.shape

```
#Perform initial analysis on final Data
final_data.shape
```

```
(111206, 9)
```

```
final_data.head(10)
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
0	1/4/2017	Sal:1	SOLANKI PLASTICS	DONA-VAI-9100	2	1,690.00	3,380.00	NaN	13,100.00
1	1/4/2017	Sal:1	SOLANKI PLASTICS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
2	1/4/2017	Sal:2	SARNESWARA TRADERS	VISHNU CHOTA WINE	500	23	11,500.00	NaN	30,990.00
3	1/4/2017	Sal:2	SARNESWARA TRADERS	LITE FOAM(1200)	6	1,620.00	9,720.00	NaN	NaN
4	1/4/2017	Sal:2	SARNESWARA TRADERS	DONA-VAI-9100	5	1,690.00	8,450.00	NaN	NaN
5	1/4/2017	Sal:2	SARNESWARA TRADERS	CLASSIC ENJOY(750)	1	1,320.00	1,320.00	NaN	NaN
6	1/4/2017	Sal:898	Lock	Vishnu 250ml	100	30	3,000.00	100	5,400.00
7	1/4/2017	Sal:898	Lock	BLACK DOG-350ML	100	26	2,600.00	100	NaN
8			khader vali late en						
9			try						

View First 10 Rows

final_data.head(10)

View Last 10 Rows

final_data.tail(10)

```
final_data.tail(10)
```

	Date	Voucher	Party	Product	Qty	Rate	Gross	Disc	Voucher Amount
111196	10/10/2019	Sal:4935	K.SRIHARI	CYCLE-BLU-10*12	1,200.00	6.6	7,920.00	NaN	34,980.00
111197	10/10/2019	Sal:4935	K.SRIHARI	16*20(100-W)	140	26	3,640.00	NaN	NaN
111198	10/10/2019	Sal:4935	K.SRIHARI	10*12 KRISHNA-BK(10	600	8.4	5,040.00	NaN	NaN
111199	10/10/2019	Sal:4935	K.SRIHARI	13*16 Bk(100)KRISHN	320	16	5,120.00	NaN	NaN
111200	10/10/2019	Sal:4935	K.SRIHARI	10*12 RK	800	8.5	6,800.00	NaN	NaN
111201	10/10/2019	Sal:4935	K.SRIHARI	13*16 WHITE RK	400	16	6,400.00	NaN	NaN
111202	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
111203	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
111204	NaN	Total	NaN	NaN	99,284.90	175,381.65	2,203,649.50	20,680.00	2,189,014.50
111205	NaN	Total	NaN	NaN	2,710,193.00	5,519,888.40	53,360,791.40	672,984.00	52,830,224.40

Check for Missing Values

final_data.isna().sum()

```
final_data.isna().sum()
```

	0
Date	12591
Voucher	12557
Party	40
Product	12591
Qty	12557
Rate	12558
Gross	12558
Disc	105609
Voucher Amount	83646

dtype: int64

```
#how to convert object into numerical
final_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 111206 entries, 0 to 111205
Data columns (total 9 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   Date            98615 non-null   object
 1   Voucher         98649 non-null   object
 2   Party           111166 non-null  object
 3   Product         98615 non-null   object
 4   Qty             98649 non-null   object
 5   Rate            98648 non-null   object
 6   Gross           98648 non-null   object
 7   Disc            5597 non-null    object
 8   Voucher Amount  27560 non-null   object
dtypes: object(9)
memory usage: 7.6+ MB
```

#what are the procedure we will follow to work on null values

Working with Null Values

Steps to Handle Missing Data

1. **Ask the client** for the missing values
2. **Drop null value records** (⚠️ leads to data loss)
3. **Fill missing values** using statistical methods:
    - Mean
    - Median
    - Mode

#how to convert object into numerical

final_data.info()

```
final_data.columns
```

```
Index(['Date', 'Voucher', 'Party', 'Product', 'Qty', 'Rate', 'Gross', 'Disc',
       'Voucher Amount'],
      dtype='object')
```

final_data.dtypes

```
final_data.dtypes
```

	0
Date	object
Voucher	object
Party	object
Product	object
Qty	object
Rate	object
Gross	object
Disc	object
Voucher Amount	object

dtype: object

ins=pd.read_csv("/home/insurance.csv")

ins

```
ins=pd.read_csv("/home/insurance.csv")
ins
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...	...	...	...	...	...	...	...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.9	0	no	southeast	1629.83
1336	21	female	25.8	0	no	southwest	2007.95
1337	61	female	29.1	0	yes	northwest	29141.36

1338 rows × 7 columns

#each region wise total expenses

#gender wise avg b mi expenses

#each region wise, each children and smoker class wise total expenses

ins.groupby(by='region')[['expenses']].sum().sort_values(by='expenses',ascending=False)

```
#each region wise total expenses
#gender wise avg b mi expenses
#each region wise, each children and smoker class wise total expenses


ins.groupby(by='region')[['expenses']].sum().sort_values(by='expenses',ascending=False)
```

region	expenses
southeast	5363689.80
northeast	4343668.64
northwest	4035711.93
southwest	4012754.82

ins.groupby(by='sex')[['bmi','expenses']].mean()

```
ins.groupby(by='sex')[['bmi','expenses']].mean()
```

sex	bmi	expenses
female	30.379758	12569.578897
male	30.945266	13956.751420

#each region wise, each children and smoker class wise total expenses

ins.groupby(by=['region' ,'children' ,'smoker'])[['expenses']].sum()

```
#each region wise, each children and smoker class wise total expenses
ins.groupby(by=['region' ,'children' ,'smoker'])[['expenses']].sum()
```

			expenses
region	children	smoker	
northeast	0	no	976747.39
		yes	732342.66
	1	no	544403.10
		yes	711482.80
	2	no	490110.44
		yes	204262.33
	3	no	221947.50
		yes	340039.14
	4	no	101396.36
	5	no	20936.92
northwest	0	no	814816.76
		yes	680000.21
	1	no	518805.68