

```
#importing pandas library
import pandas as pd

#import the required file
df=pd.read_excel("/content/train.xlsx")

#to known what are the cols in your dataset
df.columns

Index(['passenger', 'ID', 'survival', 'p class', 'Name', 'Sex', 'Age', 'Parch',
      'Tiket', 'Cabin', 'Embar'],
      dtype='object')

#top 5 rows in my database
df.head(9)
```



	passenger	ID	survival	p class	Name	Sex	Age	Parch	Tiket	Cabin	Embar
0	0	1	0	3	James	male	30	0	AI5211	NaN	s
1	1	2	1	1	Ali	male	29	0	pc17.2	C85	c
2	2	3	1	3	Musa	male	22	0	st023	NaN	s
3	3	4	1	1	Usman	male	32	0	11.38	C123	s
4	4	5	0	3	Jessy	male	25	0	3734	NaN	s
5	5	6	0	3	Zara	female	39	0	3308	NaN	q
6	6	7	0	1	Amina	female	19	3	17463	E46	s
7	7	8	1	3	Andrew	male	22	0	349905	NaN	s

Next steps:

Generate code with df

☒ View recommended plots

```
#decribe funtion used to find the total count, mean, standard deviation, minimum value, mximumum value
#25,50,75 percentage of the dataset value
df.describe()
```

	passenger	ID	survival	p class	Age	Parch	
count	10.00000	10.00000	10.000000	10.000000	10.000000	10.000000	
mean	4.50000	5.50000	0.600000	2.300000	26.500000	0.500000	
std	3.02765	3.02765	0.516398	0.948683	6.204837	0.971825	
min	0.00000	1.00000	0.000000	1.000000	19.000000	0.000000	
25%	2.25000	3.25000	0.000000	1.250000	22.000000	0.000000	
50%	4.50000	5.50000	1.000000	3.000000	26.000000	0.000000	
75%	6.75000	7.75000	1.000000	3.000000	29.750000	0.750000	
max	9.00000	10.00000	1.000000	3.000000	39.000000	3.000000	

```
#how many nulls
print(df.isnull().sum())

passenger    0
ID            0
survival     0
p class      0
Name         0
Sex          0
Age          0
Parch        0
Tiket        0
Cabin        7
Embar        0
dtype: int64
```

```
#to remove the null values we can use fillna method
#for example:
df.Cabin.fillna("unknow")
print(df.isnull().sum())
#it is one of the method of data clearing!!!!
```

```
passenger    0
ID           0
survival     0
p class      0
Name         0
Sex          0
Age          0
Parch        0
Tiket        0
Cabin        7
Embar        0
dtype: int64
```

```
#show does the data looks like
print(df.shape)#total row x column count
print("/n")
print(df.dtypes) #each column data type
```

```
(10, 11)
/n
passenger    int64
ID           int64
survival     int64
p class      int64
Name         object
Sex          object
Age          int64
Parch        int64
Tiket        object
Cabin        object
Embar        object
dtype: object
```

```
#info gives count and datatype of it
df.info()
print("-"*40)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  -
0   passenger   10 non-null    int64
1   ID          10 non-null    int64
2   survival    10 non-null    int64
3   p class     10 non-null    int64
4   Name        10 non-null    object
5   Sex         10 non-null    object
6   Age         10 non-null    int64
7   Parch       10 non-null    int64
8   Tiket       10 non-null    object
9   Cabin       3 non-null     object
10  Embar       10 non-null    object
dtypes: int64(6), object(5)
memory usage: 1008.0+ bytes
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  -
0   passenger   10 non-null    int64
1   ID          10 non-null    int64
2   survival    10 non-null    int64
3   p class     10 non-null    int64
4   Name        10 non-null    object
5   Sex         10 non-null    object
6   Age         10 non-null    int64
7   Parch       10 non-null    int64
8   Tiket       10 non-null    object
9   Cabin       3 non-null     object
10  Embar       10 non-null    object
```

```
dtypes: int64(6), object(5)  
memory usage: 1008.0+ bytes
```