

PANDAS IN A SNAP

```
import pandas as pd
# LOAD THE CSV FILE INTO A DATAFRAME. A DATAFRAME IS THE SQL EQUIVALENT OF A TABLE
df=pd.read_csv("titanic.csv")
# THIS WILL PRINT THE FIRST TOP 10 ROWS OF THE DATAFRAME
df.head(10)
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500
5	0	3	Mr. James Moran	male	27.0	0	0	8.4583
6	0	1	Mr. Timothy J McCarthy	male	54.0	0	0	51.8625
7	0	3	Master. Gosta Leonard Palsson	male	2.0	3	1	21.0750
8	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27.0	0	2	11.1333
9	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14.0	1	0	30.0708

NEW TASK→2

```
# A DATAFRAME CONSISTS OF SERIES WHICH ARE THE EQUIVALENT OF FIELDS IN SQL TABLE i.e. Name, Sex, Age.
# TO PRINT ONE SERIES
df["Name"]
```

```
0      Mr. Owen Harris Braund
1  Mrs. John Bradley (Florence Briggs Thayer) Cum...
2      Miss. Laina Heikkinen
3  Mrs. Jacques Heath (Lily May Peel) Futrelle
4      Mr. William Henry Allen
...
882      Rev. Juozas Montvila
883  Miss. Margaret Edith Graham
884  Miss. Catherine Helen Johnston
885      Mr. Karl Howell Behr
886      Mr. Patrick Dooley
Name: Name, Length: 887, dtype: object
```

NEW TASK→3

```
# TO PRINT MULTIPLE SERIES YOU PASS A LIST OF THE SERIES INTO THE DATAFRAME
df[['Name','Sex']]
```

	Name	Sex
0	Mr. Owen Harris Braund	male
1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female
2	Miss. Laina Heikkinen	female
3	Mrs. Jacques Heath (Lily May Peel) Futrelle	female
4	Mr. William Henry Allen	male
...
882	Rev. Juozas Montvila	male
883	Miss. Margaret Edith Graham	female
884	Miss. Catherine Helen Johnston	female
885	Mr. Karl Howell Behr	male
886	Mr. Patrick Dooley	male

887 rows × 2 columns

NEW TASK→4

```
# RETRIEVE A SINGLE ROW OF DATA
df.loc[4]
```

```
Survived          0
Pclass            3
Name              Mr. William Henry Allen
Sex               male
Age              35.0
Siblings/Spouses Aboard  0
Parents/Children Aboard  0
Fare              8.05
Name: 4, dtype: object
```

NEW TASK→5

```
# RETRIEVE ALL ROWS BASED ON A CLAUSE
df[(df["Age"]>= 10) & (df["Age"]<=30)]
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
5	0	3	Mr. James Moran	male	27.0	0	0	8.4583
8	1	3	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	female	27.0	0	2	11.1333
9	1	2	Mrs. Nicholas (Adele Achem) Nasser	female	14.0	1	0	30.0708
...
879	0	2	Mr. Frederick James Banfield	male	28.0	0	0	10.5000
880	0	3	Mr. Henry Jr Sutehall	male	25.0	0	0	7.0500
882	0	2	Rev. Juozas Montvila	male	27.0	0	0	13.0000
883	1	1	Miss. Margaret Edith Graham	female	19.0	0	0	30.0000
885	1	1	Mr. Karl Howell Behr	male	26.0	0	0	30.0000

454 rows × 8 columns

NEW TASK→6

```
# RETRIEVE ALL PASSENGERS NOT 30 OR 35 (USING A TILDE ~)
# ALTERNATIVELY
# df[(df["Age"]!="30") | (df["Age"]!="35")]
df[~df["Age"].isin(["30","35"])]
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500
...
882	0	2	Rev. Juozas Montvila	male	27.0	0	0	13.0000
883	1	1	Miss. Margaret Edith Graham	female	19.0	0	0	30.0000
884	0	3	Miss. Catherine Helen Johnston	female	7.0	1	2	23.4500
885	1	1	Mr. Karl Howell Behr	male	26.0	0	0	30.0000
886	0	3	Mr. Patrick Dooley	male	32.0	0	0	7.7500

887 rows × 8 columns

NEW TASK→7

```
# WHAT WOULD BE EACH PASSENGER'S AGE DEVIATION FROM THE AVERAGE AGE IN A NEW FIELD?
df["AGE DEVIATION"] = df["Age"] - df["Age"].mean()
df
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	AGE DEVIATION
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500	-7.471443
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833	8.528557
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250	-3.471443
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000	5.528557
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500	5.528557
...
882	0	2	Rev. Juozas Montvila	male	27.0	0	0	13.0000	-2.471443
883	1	1	Miss. Margaret Edith Graham	female	19.0	0	0	30.0000	-10.471443
884	0	3	Miss. Catherine Helen Johnston	female	7.0	1	2	23.4500	-22.471443
885	1	1	Mr. Karl Howell Behr	male	26.0	0	0	30.0000	-3.471443
886	0	3	Mr. Patrick Dooley	male	32.0	0	0	7.7500	2.528557

887 rows × 9 columns

NEW TASK→8

```
# RESET THE INDEX OF A FILTERED SUBSET
# RUN ALL CODES TO COMPARE RESULTS
# newdf = df[df["Age"]==30]
# newdf = df[df["Age"]==30].reset_index()
newdf = df[df["Age"]==30].reset_index(drop=True)
newdf.head(10)
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	AGE DEVIATION
0	0	3	Mr. William John Rogers	male	30.0	0	0	8.050	0.528557
1	1	3	Miss. Elizabeth Dowdell	female	30.0	0	0	12.475	0.528557
2	0	3	Mr. Harry Corn	male	30.0	0	0	8.050	0.528557
3	0	2	Mr. Reginald Hale	male	30.0	0	0	13.000	0.528557
4	0	2	Mr. Hans Kristensen Givard	male	30.0	0	0	13.000	0.528557
5	0	2	Mr. Walter Harris	male	30.0	0	0	10.500	0.528557
6	0	3	Mr. Sleiman Attalah	male	30.0	0	0	7.225	0.528557
7	0	3	Mr. William Arthur Lobb	male	30.0	1	0	16.100	0.528557
8	1	1	Miss. Gladys Cherry	female	30.0	0	0	86.500	0.528557
9	1	3	Mr. Theodore de Mulder	male	30.0	0	0	9.500	0.528557

NEW TASK→9

```
# SORTING DATA AS IN SQL ORDER BY Sex AND THEN MY Age in ascending order
df.sort_values(by=["Sex","Age"], ascending=True)
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	AGE DEVIATION
466	1	3	Miss. Helene Barbara Baclini	female	0.75	2	1	19.2583	-28.721443
641	1	3	Miss. Eugenie Baclini	female	0.75	2	1	19.2583	-28.721443
171	1	3	Miss. Eleanor Ileen Johnson	female	1.00	1	1	11.1333	-28.471443
379	1	3	Miss. Maria Nakid	female	1.00	0	2	15.7417	-28.471443
118	0	3	Miss. Ellis Anna Maria Andersson	female	2.00	4	2	31.2750	-27.471443
...
115	0	3	Mr. Patrick Connors	male	70.50	0	0	7.7500	41.028557
95	0	1	Mr. George B Goldschmidt	male	71.00	0	0	34.6542	41.528557
490	0	1	Mr. Ramon Artagaveytia	male	71.00	0	0	49.5042	41.528557
847	0	3	Mr. Johan Svensson	male	74.00	0	0	7.7750	44.528557
627	1	1	Mr. Algernon Henry Wilson Barkworth	male	80.00	0	0	30.0000	50.528557

887 rows × 9 columns

NEW TASK→10

```
# RETRIEVE TOP 10 NAMES WITH JOHN. MIND THE SPACE AFTER JOHN TO EXCLUDE JOHNSON
df[df["Name"].str.contains("John ")]
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	AGE DEVIATION
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833	8.528557
41	0	2	Mrs. William John Robert (Dorothy Ann Wonnacot...	female	27.0	1	0	21.0000	-2.471443
44	0	3	Mr. William John Rogers	male	30.0	0	0	8.0500	0.528557
97	1	2	Mrs. John T (Ada Julia Bone) Doling	female	34.0	0	1	23.0000	4.528557
111	0	3	Mr. David John Barton	male	22.0	0	0	8.0500	-7.471443
116	0	2	Mr. William John Robert Turpin	male	29.0	1	0	21.0000	-0.471443
159	0	3	Mr. John Hatfield Cribb	male	44.0	0	1	16.1000	14.528557
161	0	3	Mr. John Viktor Bengtsson	male	26.0	0	0	7.7750	-3.471443
164	1	3	Master, Frank John William Goldsmith	male	9.0	0	2	20.5250	-20.471443
167	0	1	Mr. John D Baumann	male	60.0	0	0	25.9250	30.528557

NEW TASK→11

```
# EXPORT TO A NEW .CSV FILE
df.to_csv("titanic_new_info.csv")
```

```
# MATRIX RELOAD
df=pd.read_csv("titanic.csv")
df.dtypes
```

```
Survived          int64
Pclass            int64
Name              object
Sex               object
Age              float64
Siblings/Spouses Aboard  int64
Parents/Children Aboard  int64
Fare              float64
dtype: object
```

NEW TASK→12

```
# CREATE A DATAFRAME BASED ON THE TITANIC FRAME
person_survived = {"Survived":[1],
                   "Pclass":[1],
                   "Name":["Mr. Leo Miaris"],
                   "Sex":["male"],
                   "Age":[78.0],
                   "Siblings/Spouses Aboard":[0],
                   "Parents/Children Aboard":[0],
                   "Fare":[100.0]}
person_to_add = pd.DataFrame(columns=df.columns, data=person_survived)
person_to_add
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	1	1	Mr. Leo Miaris	male	78.0	0	0	100.0

NEW TASK→13

```
# CREATE A DATAFRAME FROM SCRATCH
lookup_data = {"Pclass":[1,2,3],"Total cabin amount":[371,105,84]}
additional_info = pd.DataFrame(columns = ['Pclass','Total cabin amount'], data=lookup_data)
additional_info
```

	Pclass	Total cabin amount
0	1	371
1	2	105
2	3	84

NEW TASK→14

```
# GET ALL THE DATAFRAME STATISTICS FOR DATA WRANGLING ETC
df.describe()
```

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
count	887.000000	887.000000	887.000000	887.000000	887.000000	887.000000
mean	0.385569	2.305524	29.471443	0.525366	0.383315	32.30542
std	0.487004	0.836662	14.121908	1.104669	0.807466	49.78204
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	0.000000	2.000000	20.250000	0.000000	0.000000	7.92500
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.13750
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

NEW TASK→16

```
# AGGREGATE BASED ON SERIES
df["Age"].value_counts()
```

```
22.00    39
28.00    37
18.00    36
21.00    34
24.00    34
..
0.92     1
23.50     1
36.50     1
55.50     1
74.00     1
Name: Age, Length: 89, dtype: int64
```

NEW TASK→17

```
# GROUP BY AGGREGATION
df.groupby("Sex")["Age"].mean()
```

```
Sex
female    27.719745
male      30.431361
Name: Age, dtype: float64
```

NEW TASK→18

```
# PLOT DATA
import matplotlib
df["Age"].hist(figsize=(10,5))
```

<AxesSubplot:>



