Homework 07

⚠️Before you start ⚠️

Duplicate this Jupyter Notebook in your `week-08` *folder (right-click -> Duplicate) and then add your last name to the beginning of it (ie.* `blevins-hw-07.ipynb` *- otherwise you risk having all your work overwritten when you try to sync your GitHub repository with your instructor's repository.*

We're going to be practing using the Pandas library to explore another dataset: a famouse collection of information about some passengers on board the *Titanic*. To find out more information about this dataset look at the data dictionary on this page: https://www.kaggle.com/c/titanic/data#:~:text=should%20look%20like.-,data%20dictionary,-Variable

Import the pandas library.

In [6]:

```
#Your Code Here
import pandas as pd
```

Read in the CSV file.

In [8]:

```
#Your Code Here

titanic_df = pd.read_csv('titanic.csv',encoding='utf-8')
```

Display the first 12 rows of your dataset.

In [10]:

```
titanic_df[:11]
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549

What are the different data types contained in each column?

In [12]:

```
#Your Code Here

titanic_df.dtypes

#891 rows = 891 passengers?
```

Out[12]:

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In your own words, what is the difference in the data types for `Survived` vs. `Age` columns?

The 'Survived' column is more of a binary yes or no answer using integers. According to the data dictionary a 0 in the 'survived column means the passenger did not survive while a 1 means they did. The 'age' column is describing someone's age but as a float because that's how the data captures babies under a year old.

Use the `.isna()` or `.notna()` methods in conjunction with a filter to only select rows from your dataframe consisting of passengers for which we have information about the cabin they were in.

In [15]:

```
#Your Code Here
titanic_df[titanic_df["Cabin"].notna()]
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783
...	...	...	...	...	...	...	...	...	...
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751
872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369

204 rows × 12 columns

What percentage of rows (passengers) in the dataset have information about their

cabin number?

In [17]:

```
print(204/891)
```

```
0.22895622895622897
```

In [18]:

```
titanic_df["Cabin"].count()/len(titanic_df)
```

Out[18]:

```
0.22895622895622897
```

23% of the dataset rows have information on the passenger's cabin number

Some of our columns are hard to read. **Rename the following columns:**

- The `SibSp` column contains information about whether the passenger had family on board (siblings or spouses). **Rename the column `siblings_spouses`.**
- The `Pclass` column stands for the ticket class (1st, 2nd, or 3rd). **Rename the column `ticket_class`.**

Hint: remember to change it permanently rather than temporarily.

In [21]:

```
#Your Code Here

titanic_df = titanic_df.rename(columns={'SibSp':'siblings_spouses'})
```

Which passengers bought the nine most expensive tickets?

In [23]:

```
 titanic_df.sort_values(by='Fare', ascending=False)[:9]
```

Out[23]:

	PassengerId	Survived	Pclass	Name	Sex	Age	siblings_spouses	Parch
258	259	1	1	Ward, Miss. Anna	female	35.0	0	0
737	738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1
88	89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2
341	342	1	1	Fortune, Miss. Alice Elizabeth	female	24.0	3	2
438	439	0	1	Fortune, Mr. Mark	male	64.0	1	4
311	312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2
742	743	1	1	Ryerson, Miss. Susan Parker "Suzette"	female	21.0	2	2

What was the median age of passengers on the Titanic?

In [25]:

```
titanic_df.describe()

print("median age is 28 years old")
```

```
median age is 28 years old
```

Who was the oldest passenger on the Titanic in our dataset?

In [27]:

```
titanic_df.sort_values(by='Age', ascending= False)[:1]
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	siblings_spouses	Parch
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0

Use the `groupby` function to count how many passengers bought each class of ticket.

In [29]:

```
#Your Code Here
titanic_df.groupby("Pclass").count()['Ticket']
```

Out[29]:

```
Pclass
1    216
2    184
3    491
Name: Ticket, dtype: int64
```

Use the `groupby` function to group passengers into different classes of ticket and then calculate the median age of passengers within each ticket class.

In [31]:

```
pclass_groups =titanic_df.groupby("Pclass")


pclass_groups['Age'].median()
```

Out[31]:

```
Pclass
1    37.0
2    29.0
3    24.0
Name: Age, dtype: float64
```

Use the `groupby` function to group passengers into different classes of ticket and then calculate the median ticket fare within each ticket class.

In [34]:

```
pclass_groups['Fare'].median()
```

Out[34]:

```
Pclass
1    60.2875
2    14.2500
3     8.0500
Name: Fare, dtype: float64
```

Bonus Questions

Bonus: Make the Survived column more legible. Write a function and apply it to the dataframe that changes the 0 and 1 values to "Died" and "Lived." Then display the first 10 rows to see if it worked.

Note: when changing the values in columns, you might make mistakes. That's okay! You can always reload the dataframe from the original file to start over. When trying to answer this questions, each time you run it I'm going to have you start with the "original" dataframe so that you don't have to go back to the beginning of the notebook and run all the cells again.

In [37]:

```
titanic_df=pd.read_csv('titanic.csv')

def binary_change (number):
    if number == 0:
        return "no"
    elif number == 1:
        return "yes"


titanic_df['Survived'] = titanic_df['Survived'].apply(binary_change)
```

Bonus: What percentage of people survived the Titanic?

In [39]:

```
survivor_perc = ((titanic_df[titanic_df['Survived'] == 'yes']['Name'].count(
survivor_perc= round(survivor_perc,1)
print(f'{survivor_perc}%')
```

```
38.4%
```

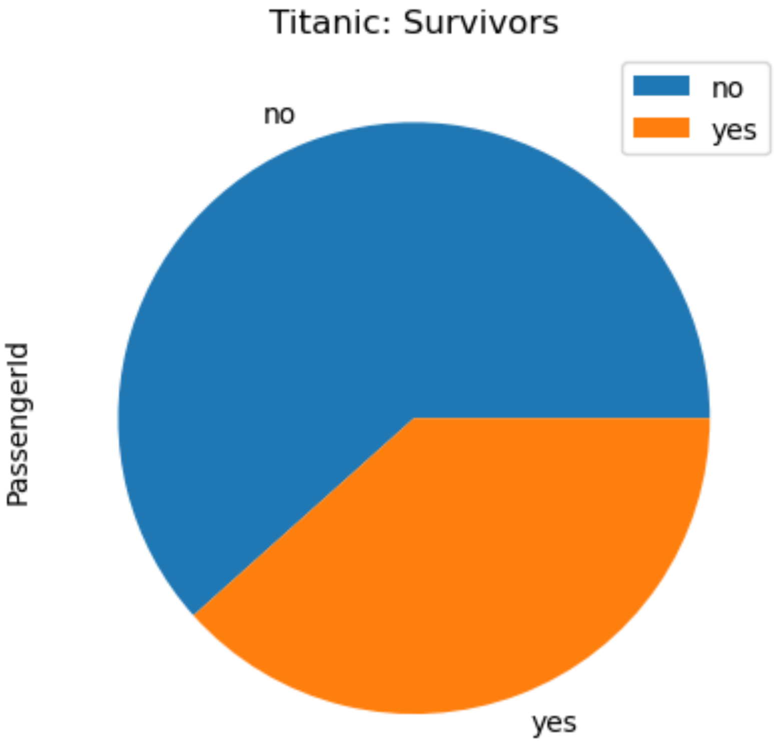
Bonus: Make a pie chart visualizing the proportion of people who survived the Titanic. Hint: use the total number of rows in the dataframe to calculate the percentage.

In [41]:

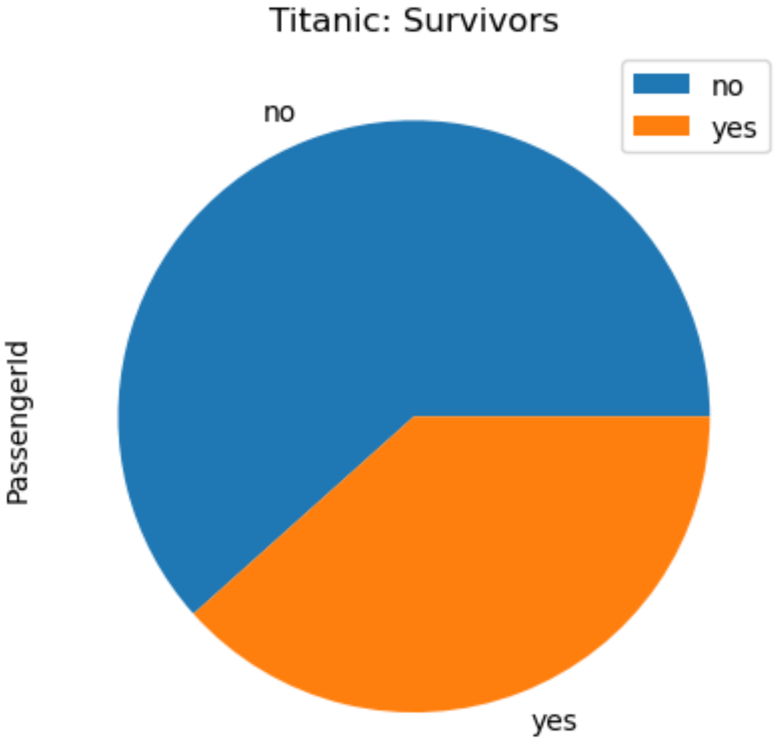
```
 ##891 rows total

survival_groups= titanic_df.groupby('Survived')['PassengerId']

survival_groups= (survival_groups.count()/891)

survival_groups.plot(kind='pie',legend= True, title="Titanic: Survivors")
```

Out[41]:

```
<Axes: title={'center': 'Titanic: Survivors'}, ylabel='PassengerId'>
```



In [42]:
```
##saving it for future reference

ax = survival_groups.plot(kind='pie',legend= True, title="Titanic: Survivors

ax.figure.savefig('titanic_survivors.png') #can also be a pdf
```



```
In [ ]:
```