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]:
```
import pandas as pd
```

Read in the CSV file.

In [8]:
```
titanic_df = pd.read_csv('titanic.csv', delimiter=",")
```

Display the first 12 rows of your dataset.

In [10]:
```
titanic_df.head(12)
```

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
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783

What are the different data types contained in each column?

In [12]:
```
titanic_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

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

Survived is an integer, whereas age is a float. Integers will have to be whole numbers (e.g. not decimals) whereas floats are decimal values. The survived is using integers to represent a true/false situation, where the only numbers should be 1 and 0.

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]:
```
cabin_filter = titanic_df['Cabin'].notna()
titanic_cabin = titanic_df[cabin_filter]
titanic_cabin
```

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]:
```
titanic_df['Cabin'].isna().value_counts(normalize=True)
```

Out[17]:
```
Cabin
True     0.771044
False    0.228956
Name: proportion, dtype: float64
```

23% of passengers in the dataset have information about their 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 [20]:
```
titanic_df = titanic_df.rename(columns={'SibSp': 'siblings_spouses'})
titanic_df = titanic_df.rename(columns={'Pclass': 'ticket_class'})
titanic_df
```

Out[20]:

	PassengerId	Survived	ticket_class	Name	Sex	Age	siblings_spouses
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0
...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0

891 rows × 12 columns

Which passengers bought the nine most expensive tickets?

In [22]:
```
titanic_df[['Name','Fare']].sort_values(by='Fare', ascending=False)[:9]
```

Out[22]:

	Name	Fare
258	Ward, Miss. Anna	512.3292
737	Lesurer, Mr. Gustave J	512.3292
679	Cardeza, Mr. Thomas Drake Martinez	512.3292
88	Fortune, Miss. Mabel Helen	263.0000
27	Fortune, Mr. Charles Alexander	263.0000
341	Fortune, Miss. Alice Elizabeth	263.0000
438	Fortune, Mr. Mark	263.0000
311	Ryerson, Miss. Emily Borie	262.3750
742	Ryerson, Miss. Susan Parker "Suzette"	262.3750

What was the median age of passengers on the Titanic?

In [24]:
```
titanic_df['Age'].median()
```

Out[24]:
```
28.0
```

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

In [26]:
```
titanic_df[['Name','Age']].sort_values(by='Age', ascending=False)[:1]
```

Out[26]:

	Name	Age
630	Barkworth, Mr. Algernon Henry Wilson	80.0

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

In [28]:
```
titanic_df.groupby('ticket_class').count()['PassengerId']
```

Out[28]:
```
ticket_class
1    216
2    184
3    491
Name: PassengerId, 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 [30]:
```
titanic_df.groupby('ticket_class')[['Age']].median()
```

Out[30]:

ticket_class	Age
1	37.0
2	29.0
3	24.0

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

In [32]:
```
titanic_df.groupby('ticket_class')[['Fare']].median()
```

Out[32]:

ticket_class	Fare
1	60.2875
2	14.2500
3	8.0500

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 [35]:
```
titanic_df=pd.read_csv('titanic.csv')

replacements = {0: 'Died', 1: 'Lived'}

titanic_df['Survived'] = titanic_df['Survived'].replace(replacements)
titanic_df
```

Out[35]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	Died	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	Lived	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	Lived	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	Lived	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	Died	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
886	887	Died	2	Montvila, Rev. Juozas	male	27.0	0	0	211536
887	888	Lived	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
888	889	Died	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607
889	890	Lived	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369
890	891	Died	3	Dooley, Mr. Patrick	male	32.0	0	0	370376

891 rows × 12 columns

Bonus: What percentage of people survived the Titanic?

In [37]:
```
survived_filter = titanic_df['Survived'] == 'Lived'
survived_df = titanic_df[survived_filter]
```

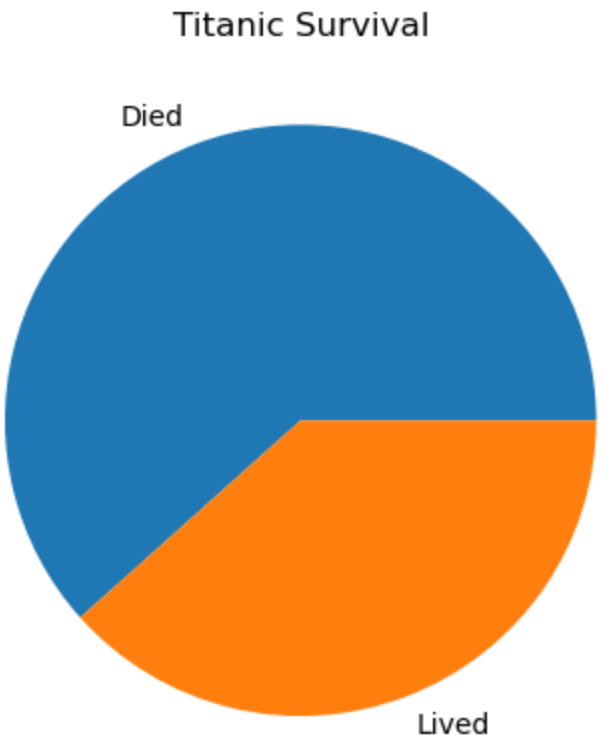
```
percent_survived = len(survived_df)/len(titanic_df)*100
print(f"{percent_survived}% of passengers survived the Titanic.")
```

```
38.38383838383838% of passengers survived the Titanic.
```

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 [39]:
```
titanic_df['Survived'].value_counts().plot(kind='pie', title = 'Titanic Surv
```

Out[39]:
```
<Axes: title={'center': 'Titanic Survival'}>
```



In [ ]:
```

```