Pandas Review Homework

Import pandas

```
In [1]: import pandas as pd
```

1. Make a data frame from a Python dictionary.

Create a Python dictionary containing

- the names of four of your friends (real or imaginary)
- their ages
- the year they started college
- their majors

```
In [3]: friends_dict = {'name': ['Kate', 'Carly', 'Talia', 'Taite'],
                        'age': [21,20,23,22],
                        'college_class': [2020, 2020, 2018, 2019],
                        'major': ['nutrition', 'neuroscience', 'religious studies', 'journalism']}
```

Make a pandas data frame from your dictionary.

```
In [4]: friends_df = pd.DataFrame(friends_dict)
```

Show your new data frame.

```
In [5]: friends_df
```

Out[5]:

	name	age	college_class	major
0	Kate	21	2020	nutrition
1	Carly	20	2020	neuroscience
2	Talia	23	2018	religious studies
3	Taite	22	2019	journalism

Fetch the ages of all your friends.

```
In [6]: friends_df['age']
```

```
Out[6]: 0    21
        1    20
        2    23
        3    22
        Name: age, dtype: int64
```

Fetch the name of your fourth friend.

```
In [7]: friends_df['name'][3]
```

```
Out[7]: 'Taite'
```

Fetch the age of your third friend.

```
In [8]: friends_df['age'][2]
```

```
Out[8]: 23
```

Compute and show the average age of your friends.

```
In [9]: friends_df['age'].mean()
```

```
Out[9]: 21.5
```

2. Find a table of data on Wikipedia and import it.

Go to Widepedia and find a table of data. It can be anything you want.

In the cell below, import the data and display it (first and last five rows).

```
In [21]: female_marathon_records = pd.read_clipboard()
         # the dataset was small enough that is displayed the whole df, so I had to do this to get
         print('This dataset didn\'t load perfectly since some people are tied')
         pd.concat([female_marathon_records.head(5),female_marathon_records.tail(5)])
```

```
This dataset didn't load perfectly since some people are tied
```

Out[21]:

	R	Time	Athlete	Date	Place	Ref
0	1	2:14:04	Brigid Kosgei (KEN)	2019.10.13	Chicago	[102]
1	2	2:14:18	Ruth Chepng'etich (KEN)	2022.10.09	Chicago	[103]
2	3	2:14:58	Amane Beriso (ETH)	2022.12.04	Valencia	[104]
3	4	2:15:25	Paula Radcliffe (GBR)	2003.04.13	London	[105]
4	5	2:15:37	Tigist Assefa (ETH)	2022.09.25	Berlin	[106][107]
18	Ashete Bekere (ETH)	2022.03.06	Tokyo	[90]	NaN	NaN
19	20	2:18:00	Rosemary Wanjiru (KEN)	2022.09.25	Berlin	[117]
20	21	2:18:03	Tigist Abayechew (ETH)	2022.09.25	Berlin	[118]
21	22	2:18:04	Joan Chelimo Melly (ROU)	2022.04.17	Seoul	[119]
22	23	2:18:05	Genzebe Dibaba (ETH)	2022.10.16	Amsterdam	[120]

3. Load the RMS titanic data and export a subset of columns

Load the titanic data, make a new `DataFrame` of the fare paid and the survival columns, and export it as a `.csv` file.

```
In [54]: titanic = pd.read_csv('data/titanic.csv')
```

Import your new `.csv` file into a new `DataFrame` and show it (first and last five rows).

In [25]:
```
titanic = pd.DataFrame(titanic)
titanic
```

Out[25]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

4. Fetch specific rows of data of the titanic data

Fetch all the second class passengers of the titanic data and put them in a new `DataFrame` and show it.

In [27]:

```
second_class = titanic[titanic['Pclass']==2].copy()
second_class
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000	NaN	S
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
20	21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	26.0000	NaN	S
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0000	D56	S
...	...	...	...	...	...	...	...	...	...	...	...	...
866	867	1	2	Duran y More, Miss. Asuncion	female	27.0	1	0	SC/PARIS 2149	13.8583	NaN	C
874	875	1	2	Abelson, Mrs. Samuel (Hannah Wizosky)	female	28.0	1	0	P/PP 3381	24.0000	NaN	C
880	881	1	2	Shelley, Mrs. William (Imanita Parrish Hall)	female	25.0	0	1	230433	26.0000	NaN	S
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.5000	NaN	S
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S

184 rows × 12 columns

Fetch all the first and third class passengers, put them in a new `DataFrame` , and show it.

In [29]:
```
second_and_third_class = titanic[(titanic['Pclass']==1)|(titanic['Pclass']==3)].copy()
second_and_third_class
```

Out[29]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

707 rows × 12 columns

5. Plot some Titanic data

First, import `matplotlib`

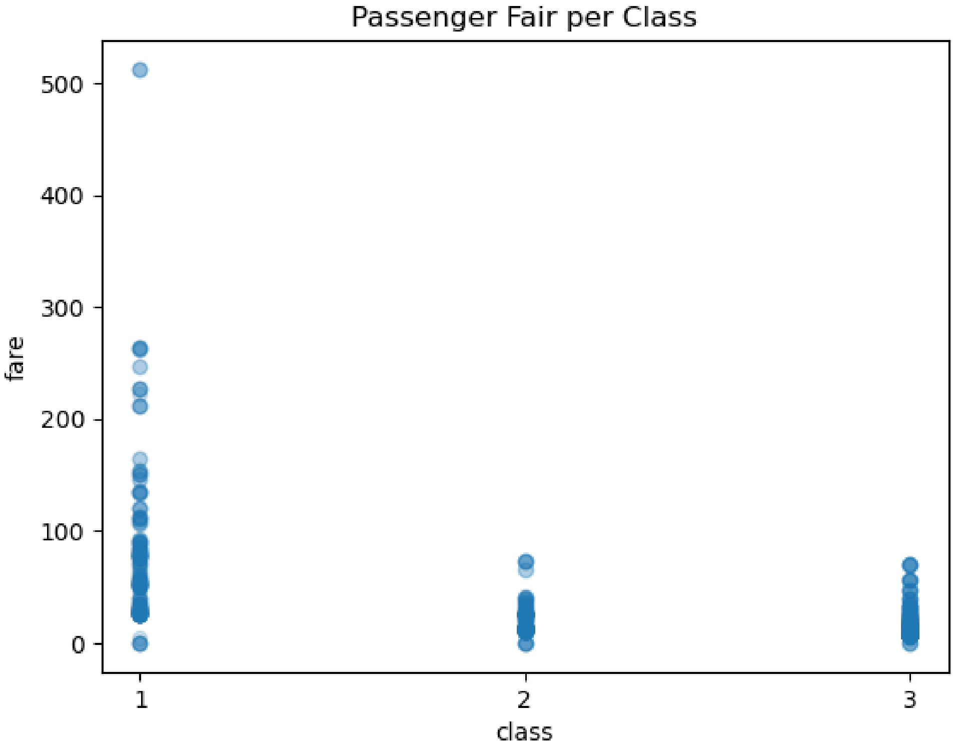
In [31]:
```
import matplotlib.pyplot as plt
```

5.a - Scatter plot

Make a scatter plot of fare vs. cabin class (seems like these should be perfectly related).

```
In [38]: plt.scatter(titanic['Pclass'], titanic['Fare'], alpha=0.2)
         plt.xticks([1,2,3])
         plt.xlabel('class')
         plt.ylabel('fare')
         plt.title('Passenger Fair per Class')
```

```
Out[38]: Text(0.5, 1.0, 'Passenger Fair per Class')
```



5.b - Distribution plot (challenging!)

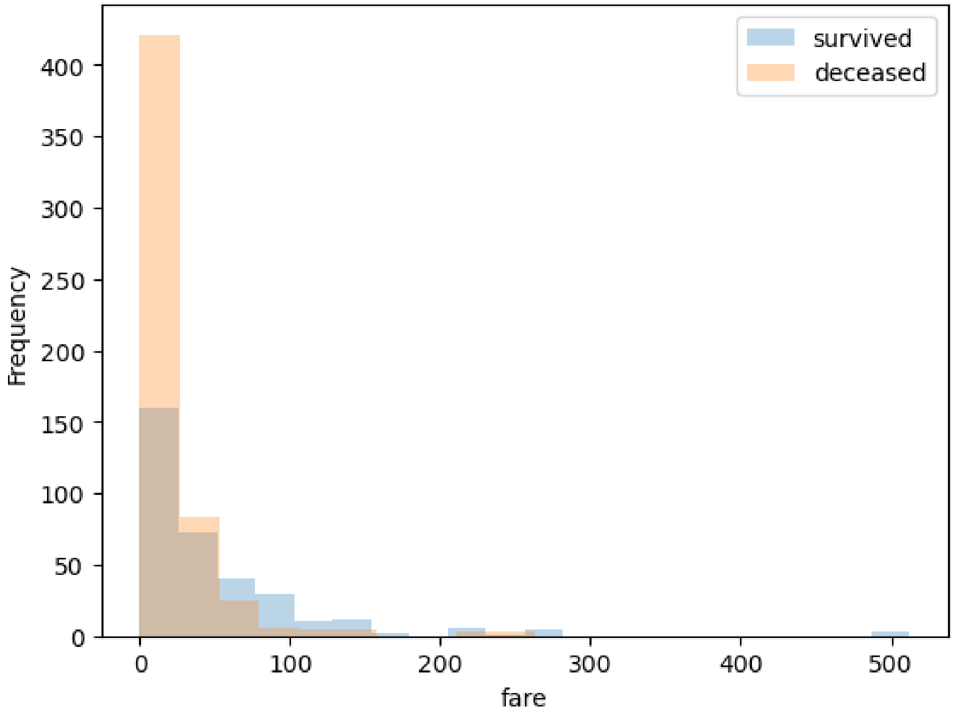
Plot the distributions of fare paid for survivors and deceased in a way that makes for a good visual comparison.

In [49]:
```
# data
survivors = titanic[titanic['Survived']==1]
deceased = titanic[titanic['Survived']==0]

#. plotting
plt.hist(survivors['Fare'], alpha=0.3, label='survived', bins=20)
plt.hist(deceased['Fare'], alpha=0.3, label='deceased')

# annotations
plt.legend()
plt.xlabel('fare')
plt.ylabel('Frequency')
```

Out[49]:
```
Text(0, 0.5, 'Frequency')
```



6. Calculate new columns

6.a - Compute total number of relatives

Create a new column in your titanic `DataFrame` quantifying the total number of relatives on board (siblings + parents – the number of siblings are in `SibSp` and the number of parents are in `Parch`).

In [56]:

```
titanic['n_relatives'] = titanic['SibSp'] + titanic['Parch']
titanic
```

Out[56]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	n_relat
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	
...	...	...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	

891 rows × 13 columns

6.b - Did a person have any relatives on board?

Add another column – a Boolean column – indicating whether each person had any relatives on board.

In [57]:
```
titanic['relatives_on_board'] = titanic['n_relatives']>0
titanic
```

Out[57]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	n_relat
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	
...	...	...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q	

891 rows × 14 columns

7. Computing descriptive statistics

7.a - Compute a mean for a column

Compute the proportion of survivors of the RMS Titanic. **Hint**: the coding of `Survival` as 0 or 1 really works to our advantage here: the proportion of survivors in any group is easily computed using a common statistical function. The 7.a section header should also give you a big clue!

In [58]:
```
titanic['Survived'].mean()
```

Out[58]:
```
0.3838383838383838
```

7.a - Compute a mean for a subset of data

Compute the proportion of survivors for the females on the RMS Titanic (you can do this in one go, or two steps, using an intermediate object containing just the female data).

In [62]:
```
titanic[titanic['Sex']=='female']['Survived'].mean()
```

Out[62]:
```
0.7420382165605095
```

7.b - Compute statistics by group

Compute the proportion of female vs. male survivors of the RMS Titanic.

In [77]:
```
females_survived = titanic[titanic['Sex']=='female']['Survived'].sum()
total_females = len(titanic[titanic['Sex']=='female']['Survived'])
percent_females = round(females_survived/total_females * 100, 2)
print(f'Out of {total_females} women aboard, {females_survived} ({percent_females}%) women

males_survived = titanic[titanic['Sex']=='male']['Survived'].sum()
total_males = len(titanic[titanic['Sex']=='male']['Survived'])
percent_males = round(males_survived/total_males * 100, 2)
print(f'Out of {total_males} men aboard, {males_survived} ({percent_males}%) men survived.

print(f'For every male that survived about {round(females_survived/males_survived)} females
```

```
Out of 314 women aboard, 233 (74.2%) women survived.
Out of 577 men aboard, 109 (18.89%) men survived.
For every male that survived about 2 females survived.
```

In [78]:
```
titanic[['Sex', 'Survived']].groupby('Sex').mean()
```

Out[78]:

	Survived
Sex	
female	0.742038
male	0.188908

Now compute the proportion of female vs. male survivors of the RMS Titanic, *along with the **s**tandard **e**rror of the **m**ean*. The **bold** type should give you a hint about the name of the method to compute the standard error. To do this, you'll need to combine the `groupby()` and `agg()` methods!

In [80]:
```
titanic[['Sex', 'Survived']].groupby('Sex').agg(['mean', 'sem'])
```

Out[80]:

	Survived	
	mean	sem
Sex		
female	0.742038	0.02473
male	0.188908	0.01631

What does this tell you about gender roles when the RMS Titanic was sunk?

Females were much more likely to survive because women and children were probably put into lifeboats first.

Compute the proportion of survivors by cabin class and their standard error.

```
In [84]: titanic[['Pclass', 'Survived']].groupby('Pclass').agg(['mean', 'sem'])
```

Out[84]:

	Survived	
	mean	sem
Pclass		
1	0.629630	0.032934
2	0.472826	0.036906
3	0.242363	0.019358

What does this tell you about socio-economic status when the RMS Titanic was sunk?

People's socioeconomic status was a strong determinant of whether or not they survived. People who were in first class were more likely to survive than those in second and third class, and people in second class were more likley to survive than those in third class.