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 [101…
```
my_dict = {'name': ['Frank','Alex', 'Henry', 'Bob'],
           'age': [20, 21, 19, 19],
           'start_college': [2022, 2021, 2022, 2023],
           'major': ['Biology', 'Psychology', 'Computer Science', 'Art']}
```

Make a pandas data frame from your dictionary.

In [102…
```
my_df = pd.DataFrame(my_dict)
```

Show your new data frame.

In [103…
```
my_df
```

Out[103]:

	name	age	start_college	major
0	Frank	20	2022	Biology
1	Alex	21	2021	Psychology
2	Henry	19	2022	Computer Science
3	Bob	19	2023	Art

Fetch the ages of all your friends.

In [104…
```
my_df['age']
```

Out[104]:
```
0    20
1    21
2    19
3    19
Name: age, dtype: int64
```

Fetch the name of your fourth friend.

In [105…
```
my_df['name'][3]
```

Out[105]:
```
'Bob'
```

Fetch the age of your third friend.

In [106…
```
my_df['age'][2]
```

Out[106]:
```
19
```

Compute and show the average age of your friends.

In [107…
```
my_df['age'].mean()
```

Out[107]:
```
19.75
```

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 [100…
```
countries_dependencies = pd.read_clipboard()
countries_dependencies
```

Out[100]:

	Country / dependency	Population	%_of_world	Date
0	World	8,088,956,000	100%	13 Feb 2024
1	China	1,409,670,000	17.40%	31 Dec 2023
2	India	1,392,329,000	17.20%	1 Jul 2023
3	United States	335,893,238	4.20%	1 Jan 2024
4	Indonesia	279,118,866	3.50%	1 Jul 2023
...	...	...	...	...
236	Christmas Island (Australia)	1,692	0%	1 Jan 2021
237	Niue (NZ)	1,689	0%	11 Nov 2022
238	Tokelau (NZ)	1,647	0%	1 Jan 2019
239	Cocos (Keeling) Islands (Australia)	593	0%	30 Jun 2020
240	Pitcairn Islands (UK)	47	0%	1 Jul 2021

241 rows × 4 columns

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 [108…
```
titanic = pd.read_csv("data/titanic.csv")
```

In [32]:
```
titanic_fs = titanic[['Fare', 'Survived']]
titanic_fs.to_csv('data/titanic_fs.csv')
```

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

In [33]:
```
fare_survived = pd.read_csv("data/titanic_fs.csv")

fare_survived
```

Out[33]:

	Unnamed: 0	Fare	Survived
0	0	7.2500	0
1	1	71.2833	1
2	2	7.9250	1
3	3	53.1000	1
4	4	8.0500	0
...	...	...	...
886	886	13.0000	0
887	887	30.0000	1
888	888	23.4500	0
889	889	30.0000	1
890	890	7.7500	0

891 rows × 3 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 [110…
```
Pclass_2 = titanic[titanic['Pclass'] == 2]
Pclass_2
```

Out[110]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	3
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	1
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	1
20	21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	2
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	1
...	...	...	...	...	...	...	...	...	...	
866	867	1	2	Duran y More, Miss. Asuncion	female	27.0	1	0	SC/PARIS 2149	1
874	875	1	2	Abelson, Mrs. Samuel (Hannah Wizosky)	female	28.0	1	0	P/PP 3381	2
880	881	1	2	Shelley, Mrs. William (Imanita Parrish Hall)	female	25.0	0	1	230433	2
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	1
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	1

184 rows × 12 columns

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

In [45]:
```
Pclass_1and3 = titanic[titanic['Pclass'] != 2]
Pclass_1and3
```

Out[45]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05
...	...	...	...	...	...	...	...	...	...	
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.12
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75

707 rows × 12 columns

5. Plot some Titanic data

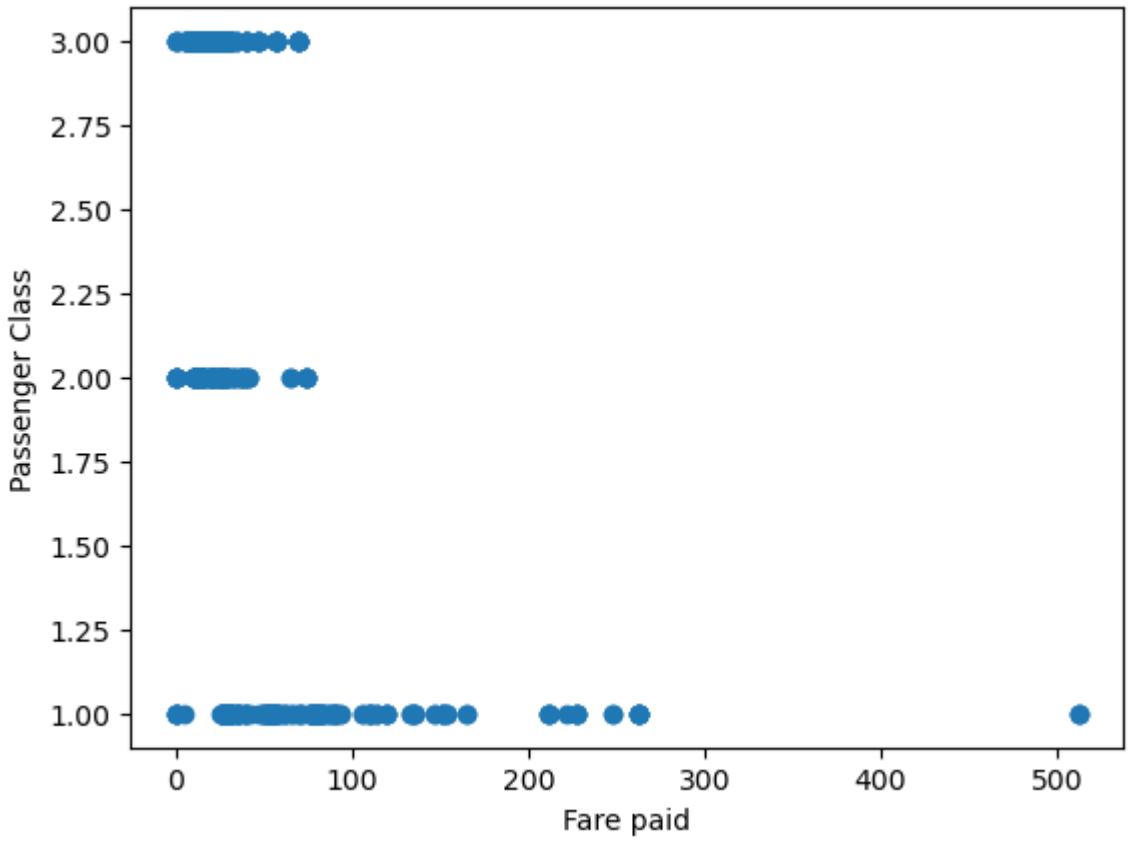
First, import `matplotlib`

In [5]:
```
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 [112…
```
plt.scatter(x = titanic['Fare'], y = titanic['Pclass']);
plt.xlabel('Fare paid')
plt.ylabel('Passenger Class');
```



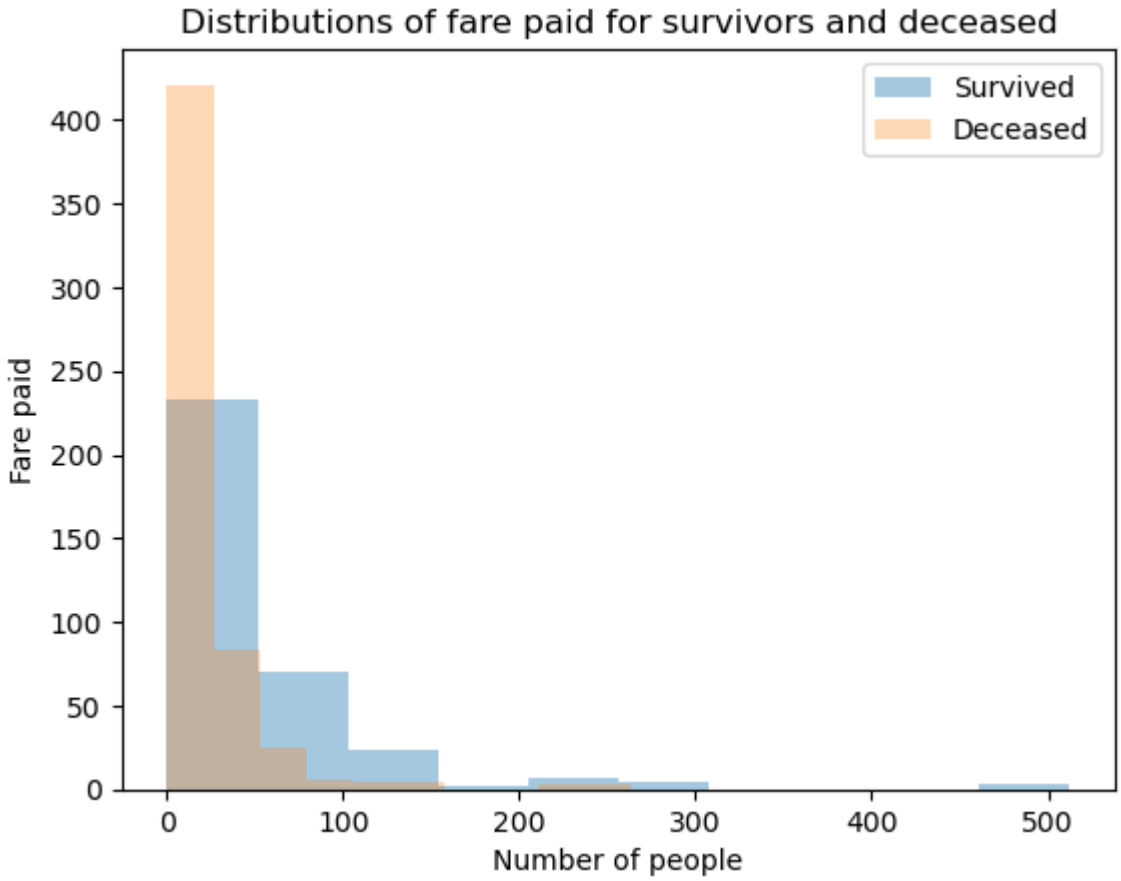
5.b - Distribution plot (challenging!)

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

In [117…
```
# filter 'Survived' to be referred to separately
survived = titanic[titanic['Survived'] == 1]
deceased = titanic[titanic['Survived'] == 0]

# plot
plt.hist(x = survived['Fare'], bins = 10, alpha = 0.4, label = 'Survived');
plt.hist(x = deceased['Fare'], bins = 10, alpha = 0.3, label = 'Deceased');

# labels and legend
plt.xlabel('Number of people')
plt.ylabel('Fare paid')
plt.title('Distributions of fare paid for survivors and deceased')
plt.legend();
```



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 [118…

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

Out[118]:

	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	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7

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 [119…

```
titanic['Y_relatives'] = titanic['n_relatives'] > 0
titanic
```

Out[119]:

	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	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7

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 [120…
```
titanic['Survived'].mean()
```

Out[120]:
```
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 [121…
```
# .loc selects all the True from the boolean titanic['Sex'] == 'female'

titanic.loc[titanic['Sex'] == 'female', 'Survived'].mean()
```

Out[121]:
```
0.7420382165605095
```

7.b - Compute statistics by group

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

In [122…
```
titanic[['Survived', 'Sex']].groupby('Sex').mean()
```

Out[122]:

	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 [123…
```
titanic_dict = {
    'Survived': ['mean', 'sem']
    }

titanic.groupby('Sex').agg(titanic_dict)
```

Out[123]:

	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?

Female passengers were probably given the priority to escape the cruise, and male passengers were more likely to stay behind to faciliate.

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

In [124…
```
titanic_dict = {
    'Survived': ['mean', 'sem']
    }

titanic.groupby('Pclass').agg(titanic_dict)
```

Out[124]:

	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?

Higher socio-economic status is associated with higher rates of survival. People of higher class were likely prioritized to evacuate the cruise.