Can A.I. save Jack from the Titanic?

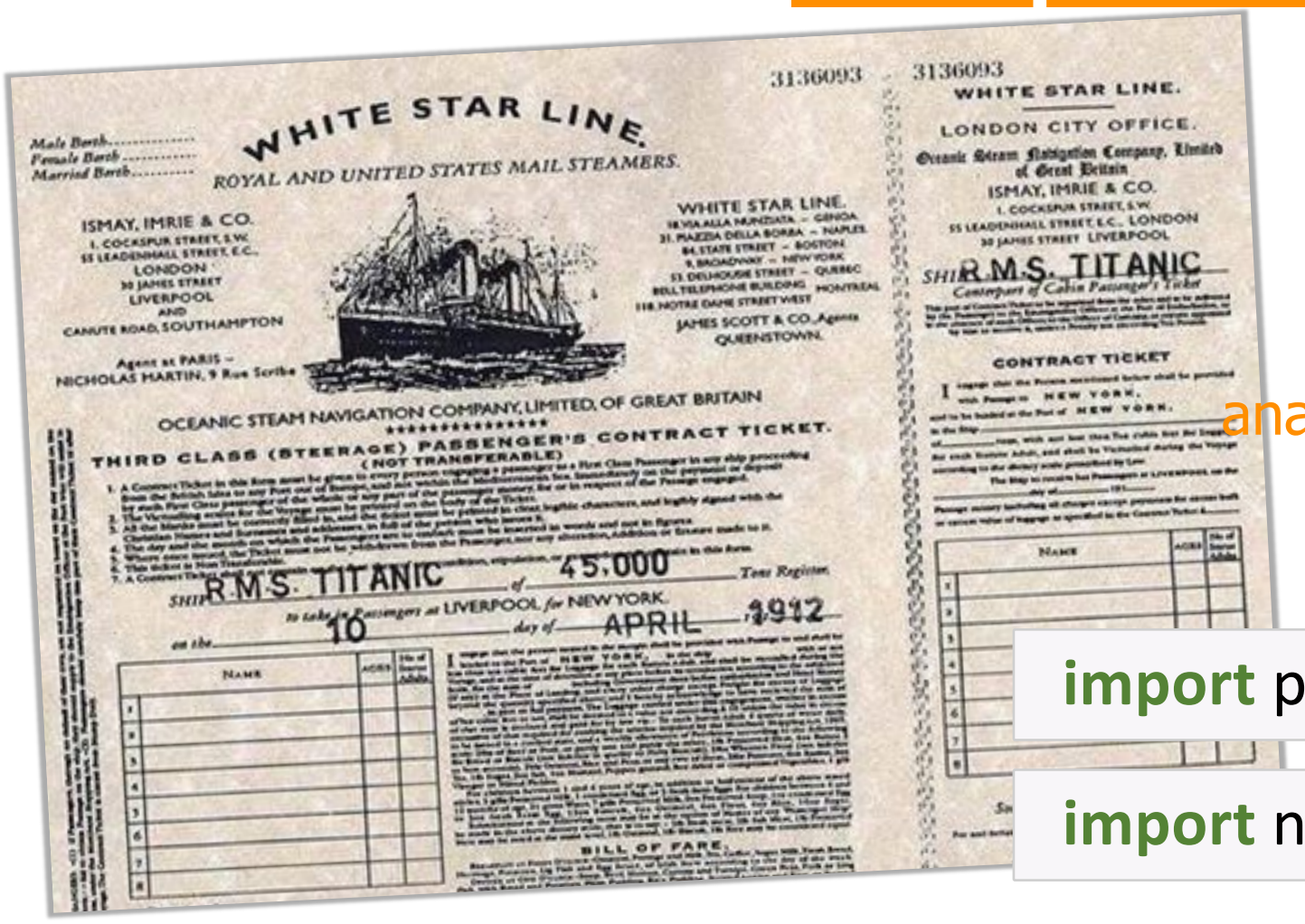
Python
Jupyter
Part 1

RMS Titanic

British passenger liner that sank in the North Atlantic Ocean in the early morning hours of April 15, 1912, after striking an iceberg during the voyage from Southampton to New York City.

Of the estimated 2,224 passengers and crew aboard, more than 1,500 died, making the sinking one of modern history's deadliest peacetime commercial marine disasters.

Import the data



First, let's import data analysis and manipulation libraries:

```
import pandas as pd
```

```
import numpy as np
```

```
# Load the dataset file into "titanic" object
```

```
titanic = pd.read_csv("titanic-1309-rows.csv")
```

```
# Let's have a look at the dataset attributes
```

```
titanic.head()
```

P. Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Brigg	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmin	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

1309 entries across **10 variables**

Survived 0 = No, 1 = Yes

Pclass Ticket category from first to third class

Fare Passenger fare

Ticket Passenger ticket number

sex

age

sibsp, **parch** number of siblings or spouses aboard, number of parents or children aboard

Cabin cabin number, **embarked** port of embarkation

```
titanic.info()
```

Clean the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
0    pclass    1309 non-null    int64
1    survived    1309 non-null    int64
2    name    1309 non-null    object
3    sex    1309 non-null    object
4    age    1046 non-null    float64
5    sibsp    1309 non-null    int64
6    parch    1309 non-null    int64
7    ticket    1309 non-null    object
8    fare    1308 non-null    float64
9    cabin    295    non-null    object
10    embarked    1307 non-null    object
11    boat    486    non-null    object
```

Here, we drop few variables such as the names, cabin and ticket numbers)

```
titanic.drop(['name','body','boat','cabin','ticket','
embarked'],axis=1,inplace=True)
```

We will fill the age and fare variables by the median value per sex type and passenger class respectively.

```
titanic['age'] = titanic.groupby('sex')['age'].apply(lambda
x: x.fillna(x.median()))
```

```
titanic[‘fare'] =
titanic.groupby(‘pclass')[‘fare'].apply(lambda x:
x.fillna(x.median()))
```

Analyse the data

```
exploratory.groupby
(['sex_is_male','pclass'])['survived'].mean()
```

sex_is_male	pclass	
0 (female)	1	0.965278
	2	0.886792
	3	0.490741
1 (male)	1	0.340782
	2	0.146199
	3	0.152130

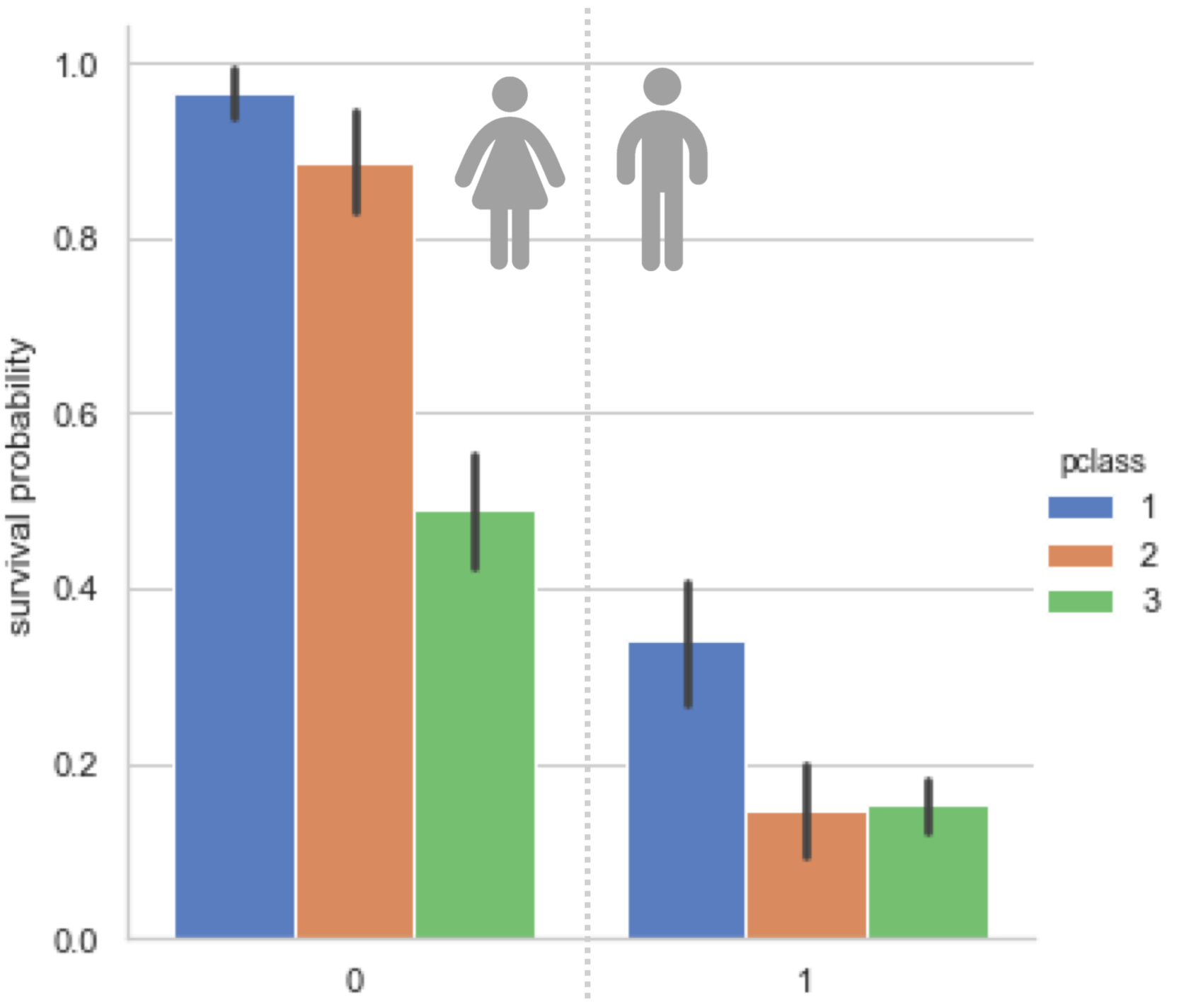
Name: survived, dtype: float64

Survival probability: for women on 1st class is: 96,5% compared to men only 34,1%

When we look at the 3rd class, the probability drops to 49,1% for women and 15,2% for men. But let's plot it!

```
import seaborn as sns
```

```
graph = sns.catplot(x="sex_is_male",
y="survived", hue="pclass", kind="bar",
palette="muted", data=titanic)
```



End of part 1

In the next sessions, we will be :

- Splitting the dataset for training and testing
- Applying Logistic Regression
- Evaluating the model
- What about Jack and Rose?
- Applying Decision Tree
- Final remarks