Quiz 2 - Data Preprocessing

```
import pandas as pd
import numpy as np
df = pd.read_excel('ds-quiz_02.xlsx')
df.shape
```

```
(891, 12)
```

```
df.head()
```

	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	female	38.0	1	0	PC 17599	7

```
df.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	3(

```
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
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Drop columns: PassengerID, Name, Ticket, Cabin

```
df.drop(['PassengerId', 'Name', 'Ticket', 'Cabin'], axis=1, inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Survived  891 non-null    int64  
 1   Pclass    891 non-null    int64  
 2   Sex       891 non-null    object 
 3   Age       714 non-null    float64
 4   SibSp     891 non-null    int64  
 5   Parch     891 non-null    int64  
 6   Fare      891 non-null    float64
 7   Embarked  889 non-null    object 
dtypes: float64(2), int64(4), object(2)
memory usage: 55.8+ KB
```

Fill `NaN` values in `age` and `fare` with the column mean

```
df['Age'] = df['Age'].fillna(df['Age'].mean())
df['Fare'] = df['Fare'].fillna(df['Fare'].mean())
df.isnull().sum()
```

```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    2
dtype: int64
```

Fill `NaN` values in `embarked` with the column mode

```
embarked_mode = df['Embarked'].mode()[0]
df['Embarked'] = df['Embarked'].fillna(embarked_mode)
df.isnull().sum()
```

```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    0
dtype: int64
```

summary statistics

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Survived	891.0	0.383838	0.486592	0.00	0.0000	0.000000	1.0	1.0000
Pclass	891.0	2.308642	0.836071	1.00	2.0000	3.000000	3.0	3.0000
Age	891.0	29.699118	13.002015	0.42	22.0000	29.699118	35.0	80.0000
SibSp	891.0	0.523008	1.102743	0.00	0.0000	0.000000	1.0	8.0000
Parch	891.0	0.381594	0.806057	0.00	0.0000	0.000000	0.0	6.0000
Fare	891.0	32.204208	49.693429	0.00	7.9104	14.454200	31.0	512.3292

Average fare by passenger class

```
df.groupby('Pclass')['Fare'].mean()
```

```
Pclass
1    84.154687
2    20.662183
3    13.675550
Name: Fare, dtype: float64
```

Average age of male and female passengers by passenger class

```
df.pivot_table(index='Pclass', columns='Sex', values='Age', aggfunc=np.mean)
```

Sex	female	male
Pclass		
1	34.141405	39.287717
2	28.748661	30.653908
3	24.068493	27.372153

Number of male and female survivors & non-survivors by passenger class

```
pd.crosstab([df['Sex'], df['Pclass']], df['Survived'], margins=True)
```

Survived		0	1	All
Sex	**Pclass**			
female	**1**	3	91	94
	2	6	70	76
	3	72	72	144
male	**1**	77	45	122
	2	91	17	108
	3	300	47	347
All		549	342	891