Task 2 - Calculate summary statistics

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [85]:
```
test=pd.read_csv("test.csv")
```

In [9]:
```
train=pd.read_csv("train.csv")
```

In [12]:
```
df1.head()
```

Out[12]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Eml
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	

In [6]:
```
df2.head()
```

Out[6]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

Check Calculate summary statistics

In [61]:
```
test.describe(include = "all")
```

Out[61]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
count	418.000000	418.000000	418	418	418.000000	418.000000	418.000000	418
unique	NaN	NaN	418	2	NaN	NaN	NaN	363
top	NaN	NaN	Kelly, Mr. James	male	NaN	NaN	NaN	PC 17608
freq	NaN	NaN	1	266	NaN	NaN	NaN	5
mean	1100.500000	2.265550	NaN	NaN	30.216507	0.447368	0.392344	NaN
std	120.810458	0.841838	NaN	NaN	12.635016	0.896760	0.981429	NaN
min	892.000000	1.000000	NaN	NaN	0.170000	0.000000	0.000000	NaN
25%	996.250000	1.000000	NaN	NaN	23.000000	0.000000	0.000000	NaN
50%	1100.500000	3.000000	NaN	NaN	30.000000	0.000000	0.000000	NaN
75%	1204.750000	3.000000	NaN	NaN	35.750000	1.000000	0.000000	NaN
max	1309.000000	3.000000	NaN	NaN	76.000000	8.000000	9.000000	NaN

In [13]:
```
train.describe()
```

Out[13]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fa
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.0000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.2042
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.6934
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.0000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.9104
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.4542
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.0000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.3292

In [18]:
```
train.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
```

Finding null values

In [15]:
```
test.isnull().sum()
```

Out[15]:
```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [17]:
```
train.isnull().sum()
```

Out[17]:
```
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
```

-- In Test table we have null values. -- Column: Age / Cabin Age: 86 null values Cabin: 327 null values

-- In train table also have null values. -- Column: Age / Cabin Age: 177 null values Cabin: 687 null values Embarked: 2 null values

Calculate summary statistics

(mean, median,mode, standard deviation) for a dataset

In [88]:
```
# Age ill fill with mean values
test.Age.fillna(30,inplace=True)
```

In [74]:
```
test.isnull().sum()
```

Out[74]:
```
PassengerId    0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Embarked       0
dtype: int64
```

In [91]:
```
test.isnull().sum()
```

Out[91]:
```
PassengerId    0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

In [72]:
```
test = test.dropna(axis='columns')
```

In [90]:
```
test.dropna(inplace=True)
```

In [ ]: