Loading the data

In [1]:

```
# importing required libraries
import pandas as pd
```

In [2]:

```
#loading the data
data = pd.read_csv('titanic_train.csv')
data.head()
```

Out[2]:

	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

Missing Values

In [3]:

```
#missing values in the data
data.isnull().sum()
```

Out[3]:

```
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
```

- Age and Cabin have a very high number of missing values
- Embarked has very low number of missing values

Deleting Data points with missing values

In [4]:

```
# Age variable without missing values treatment
data['Age'].head(6)
```

Out[4]:

```
0    22.0
1    38.0
2    26.0
3    35.0
4    35.0
5     NaN
Name: Age, dtype: float64
```

In [5]:

```
# dropping all rows with missing values

data_row_del = data.dropna(axis=0)
data_row_del['Age'].head(6)
```

Out[5]:

```
1     38.0
3     35.0
6     54.0
10     4.0
11    58.0
21    34.0
Name: Age, dtype: float64
```

- Have deleted rows, if any one column/ feature has missing values in that row

In [6]:

```
# shape before and after removing missing values

data.shape, data_row_del.shape
```

Out[6]:

```
((891, 12), (183, 12))
```

- Significant loss of information
- Only three columns had missing values

Deleting columns with missing values

In [7]:

```
## isnull with ratio

(data.isnull().sum())/891
```

Out[7]:

```
PassengerId    0.000000
Survived       0.000000
Pclass         0.000000
Name           0.000000
Sex            0.000000
Age            0.198653
SibSp          0.000000
Parch          0.000000
Ticket         0.000000
Fare           0.000000
Cabin          0.771044
Embarked       0.002245
dtype: float64
```

In [8]:

```
data.head(10)
```

Out[8]:

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

In [9]:

```
# dropping all columns with missing values

data_col_del = data.dropna(thresh = 500, axis=1)
data_col_del.head()
```

Out[9]:

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

In [10]:

```
# shape before and after removing missing values

data.shape, data_col_del.shape
```

Out[10]:

```
((891, 12), (891, 11))
```

- A better way to deal with missing values without loss of information?

Replacing with a new category/value

In [11]:

```
data['Cabin'].head()
```

Out[11]:

```
0     NaN
1     C85
2     NaN
3    C123
4     NaN
Name: Cabin, dtype: object
```

In [12]:

```
data['Cabin'].fillna(value='missing')
```

Out[12]:

```
0      missing
1          C85
2      missing
3         C123
4      missing
        ...
886    missing
887        B42
888    missing
889       C148
890    missing
Name: Cabin, Length: 891, dtype: object
```

In [13]:

```
data['Age'].fillna(value=999)
```

Out[13]:

```
0       22.0
1       38.0
2       26.0
3       35.0
4       35.0
       ...
886     27.0
887     19.0
888    999.0
889     26.0
890     32.0
Name: Age, Length: 891, dtype: float64
```

In [14]:

```
## make a copy
data_replace = data.copy()

# replace values
data_replace['Age'] = data_replace['Age'].fillna(value=999)
data_replace.isnull().sum()
```

Out[14]:

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

In [15]:

```
(data['Cabin'].isnull()).astype('int')
```

Out[15]:

```
0      1
1      0
2      1
3      0
4      1
      ..
886    1
887    0
888    1
889    0
890    1
Name: Cabin, Length: 891, dtype: int32
```

In [16]:

```
data_replace['Cabin_na'] = (data['Cabin'].isnull()).astype('int')
```

In [17]:

```
data_replace.head()
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin_na
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	0
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	1

- Similarly we can do for 'Embarked'
- Can we impute missing values with more sensible numbers?