Data Source: https://www.kaggle.com/c/titanic

IMPORT LIBRARIES AND DATASETS

```
# Import the pandas
import pandas as pd
# Import numpy
import numpy as np
# Import matplotlib.pyplot as plt
import matplotlib.pyplot as plt

# Read the data using pandas dataframe
df = pd.read_csv('train.csv')

# View the data head
df.head()
```

	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

```
df.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

PERFORM DATA VISUALIZATION

```
# Use a count plot to show the number of dead and survived
sns.countplot('survived',data=df)
plt.show()
```

```
survived_df = df[df['Survived'] == 1]
no_survived_df = df[df['Survived'] == 0]

# Count the survived and deceased
survived = len(survived_df)
print("Number of passengers who survived:", survived)
print("Percentage survived:", 1 * survived/len(df) * 100.0, "%")

deceased = len(no_survived_df)
print("Number of passengers who did not survive:", deceased)
print("Percentage did not survive:", 1 * deceased/len(df) * 100.0, "%")

total = len(df)
labels = ['Passengers who survived', 'Passengers who did not survive']
plt.bar(labels[0], survived, color='blue')
plt.bar(labels[1], deceased, color='red')
plt.title('Number of Passengers who survived vs Passengers who did not survive')
plt.show()
```

```
Number of passengers who survived: 342
Percentage survived: 38.3838383838 %
Number of passengers who did not survive: 549
Percentage did not survive: 61.6161616162 %
```

- **Do you have a higher chance of survival if you have sons or are sitting?**

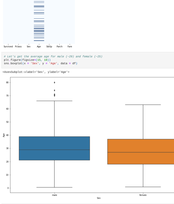
```
df.columns
```

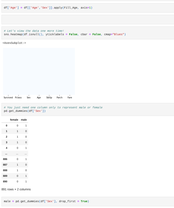
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

```
sns.countplot(x='Sex',data=df, hue='Survived')
```

```
# Use Count to indicate the number of people survived based on their sex
# If you have a female, you have a higher chance of survival compared to other parts
df['Sex'].value_counts()
pd.crosstab('Sex', 'Survived')
sns.countplot('Sex', hue = df['Survived'], data = df)
plt.show()
pd.crosstab('Sex', 'Survived').plot.bar(stacked=True)
```

```
male    577
female  314
Name: Sex, dtype: int64
```

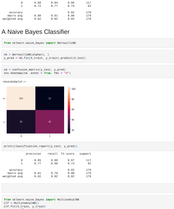





```
# Age histogram
df['Age'].hist(bins=40)
```

- **Plot the histogram for the "Fare" column**

```
df['Fare'].hist(bins=40)
```



PERFORM DATA CLEANING AND FEATURE ENGINEERING

```
# List a sample which possess or miss a cabin
df[df.Cabin.isnull() ].sample(5), df[df.Cabin.isnull() ].shape, df[df.Cabin.isnull()].shape
```

```
(     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
 ...        ...       ...     ...                                    ...     ...   ...    ...    ...     ...     ...   ...      ...
 886        887         0       2                      Montvila, Rev. Juozas    male  27.0      0      0    211536  13.0000   NaN        S
 887        888         1       1               Graham, Miss. Margaret Edith  female  19.0      0      0    112053  30.0000   B42        S
 888        889         0       3   Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1      2  W./C. 6607  23.4500   NaN        S
 889        890         1       1                          Behr, Mr. Karl Howell    male  26.0      0      0    111369  30.0000  C148        C
 890        891         0       3                            Dooley, Mr. Patrick    male  32.0      0      0    370376   7.750   NaN        Q
 [687 rows x 12 columns],
 (891, 12),
 (687, 12))
```

```
df.head()
```

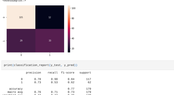
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 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

```
df
```

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

891 rows × 12 columns

```
# Use a count plot to show the number of dead and survived passengers based on their class
df.Cabin[df.Pclass == 1].value_counts()
sns.countplot(x='Pclass', hue = df.Cabin.isnull(), data = df)
plt.show()
```




```
# Use a count plot to show gender and survival
sns.countplot(x='Sex', hue = df.Cabin.isnull(), data = df)
plt.show()
```

```
def full_age(cols):
    age = cols[0]
    sex = cols[1]
    if pd.isnull(age):
        if sex == 'male':
            return 29
        else:
            return 25
    else:
        return age
```

```
df['Age'] = df[['Age','Sex']].apply(full_age, axis=1)
```

```
# Use a view the data one more time
sns.countplot(x='Sex', hue = df.Cabin.isnull(), data = df)
plt.show()
```



```
# Use list dummy variables
sex = pd.get_dummies(df['Sex'], drop_first = True)
```

```
# Concat the 3 drop one categorical and one
df['Drop_Cabin'] = df.Cabin.isnull()*1
```

```
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Drop_Cabin
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	1
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	1
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q	1

891 rows × 13 columns

```
# Use Get's add the encoded column with sport
df = pd.concat([df, sex], axis= 1)
```

```
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Drop_Cabin	male
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	1	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	0	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	0	0
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S	1	1
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	0	0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S	1	0
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	0	1
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q	1	1

891 rows × 14 columns

Scale the features

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df[['Age','Fare']] = scaler.fit_transform(df[['Age','Fare']])
df.head()
```

TRAIN LOGISTIC REGRESSION CLASSIFIER MODEL

```
# Utilizing capsulize implementation in the following cell
# Import sklearn.linear_model import LogisticRegression
lr = LogisticRegression(fit_intercept=True,solver='lbfgs',max_iter=1000,fit_intercept, w_trace)
```

ASSESS TRAINED MODEL PERFORMANCE

```
y_pred = clf.predict(X_test)
y_pred
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       
```