PRN - 240340128031

Q1

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

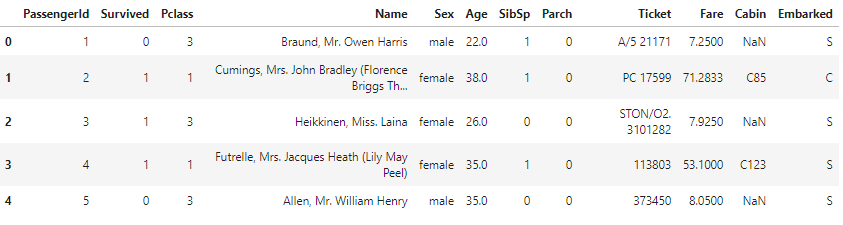
import seaborn as sns

# loading DataSet

titanic = pd.read\_csv("Titanic-Dataset.csv")

titanic.head(5)

1\_1

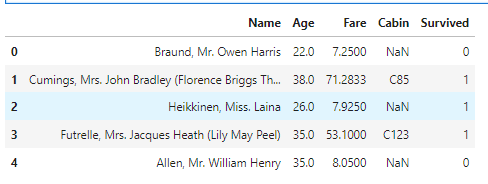


#Extract columns Name, Age, Fare, Cabin, Survived

titanic = titanic[["Name","Age","Fare","Cabin","Survived"]]

titanic.head(5)

1\_2

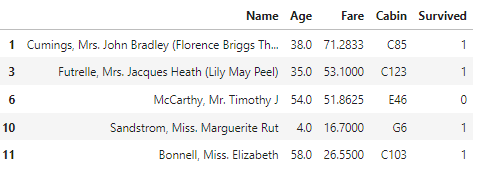


#Remove missing values from datset

titanic.dropna(inplace=True)

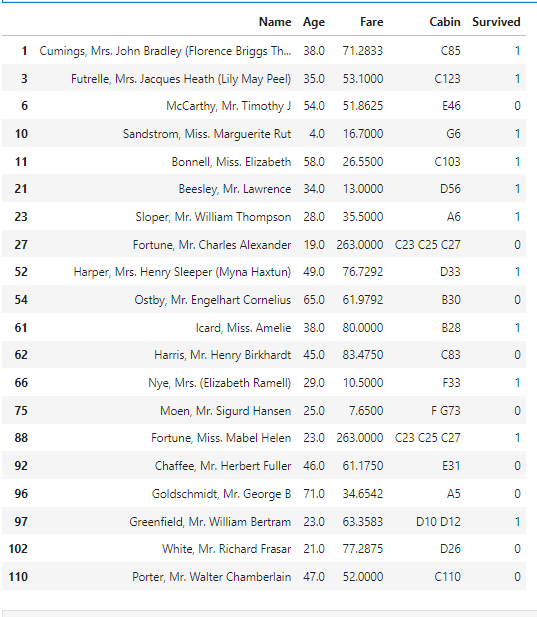
titanic.head(5)

1\_3



#Print top 20 rowss fro columns Name, Age, Fare, Cabin, Survived

titanic.head(20)

1\_4

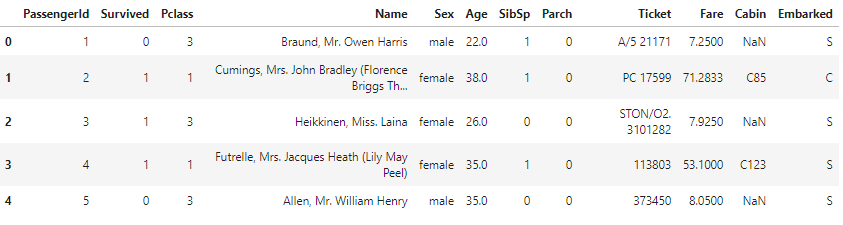
Q2

#load dataset again

titanic = pd.read\_csv("Titanic-Dataset.csv")

titanic.head(5)

2\_1



# Compute mean of Fare column

mean = titanic["Fare"].mean()

Mean

2\_2



#replace Nan values in Fare column with mean

titanic["Fare"].fillna(mean,inplace=True)

# Check if Nan values are replaced

titanic["Fare"].isna().sum()

2\_3



# Calculate 0.25 quartile, 0.75 quartile and standard deviation

q1 = titanic['Fare'].quantile(0.25)

q3 = titanic['Fare'].quantile(0.75)

iqr = q3-q1

std = titanic['Fare'].std()

# Define Lower and Upper bound

lower = q1 - 1.5\*iqr

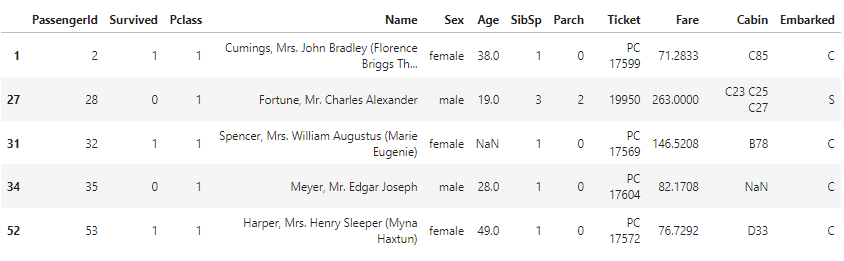
upper = q3 + 1.5\*iqr

# Check values lower than lower bound and higher than upper bound to find outliers

outliers = titanic[(titanic["Fare"]<lower)|(titanic["Fare"]>upper)]

outliers.head()

2\_4



# Boxplot for Fare values

plt.figure(figsize=(12,8))

sns.boxplot(titanic["Fare"])

plt.axhline(q1,color = "g",lw = 0.5,label = 'q1')

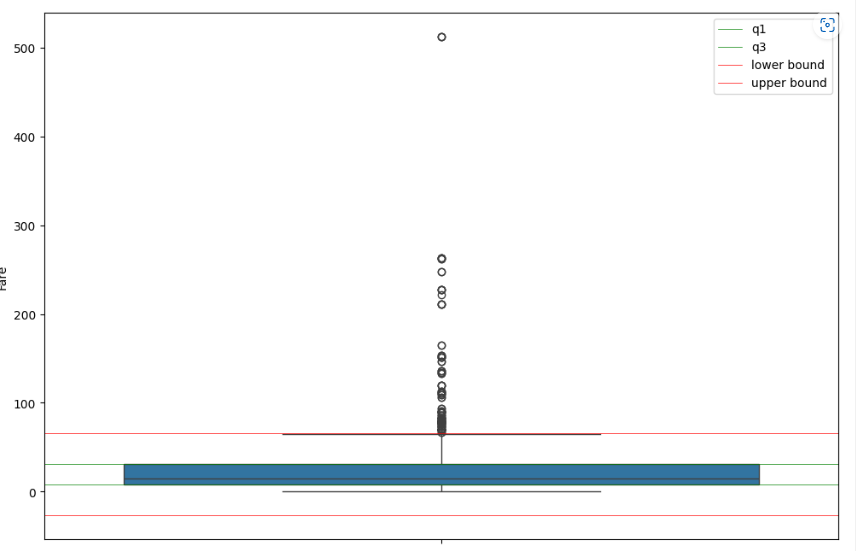
plt.axhline(q3,color = "g",lw = 0.5,label = 'q3')

plt.axhline(lower,color = "red",lw = 0.5,label = 'lower bound')

plt.axhline(upper,color = "red",lw = 0.5,label = 'upper bound')

plt.legend()

2\_5



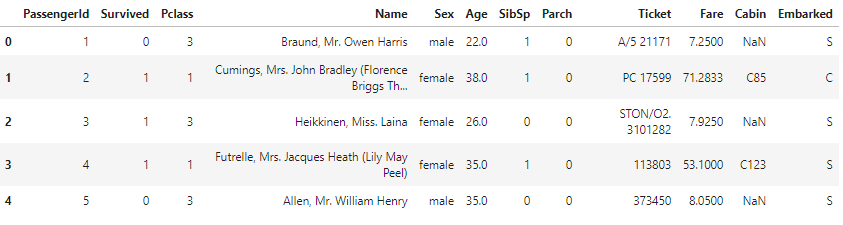
Q3

#load dataset again

titanic = pd.read\_csv("Titanic-Dataset.csv")

titanic.head(5)

3\_1



# Column with max missing values

col = titanic.isna().sum().idxmax()

print("Column with max missing values - ", col)

3\_2



# Means of the Pclass values

mean1 = (titanic["Pclass"] == 1).mean()

mean2 = (titanic["Pclass"] == 2).mean()

mean3 = (titanic["Pclass"] == 3).mean()

#Dictionary to map 'Pclass' to mean

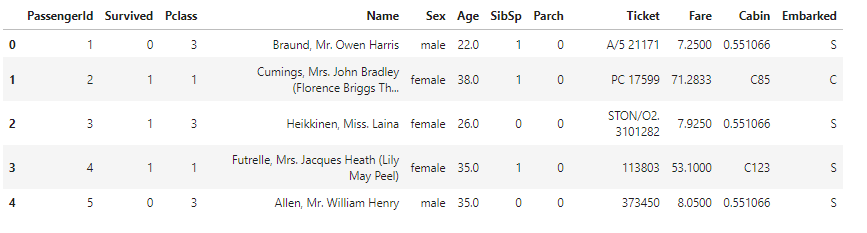
mean\_dict = {1: mean1, 2: mean2, 3: mean3}

# Replace missing values in 'max\_missing\_data' with mean of respective 'Pclass'

titanic[col] = titanic.apply(lambda row: mean\_dict[row['Pclass']] if pd.isnull(row[col]) else row[col],axis=1)

titanic.head()

3\_3



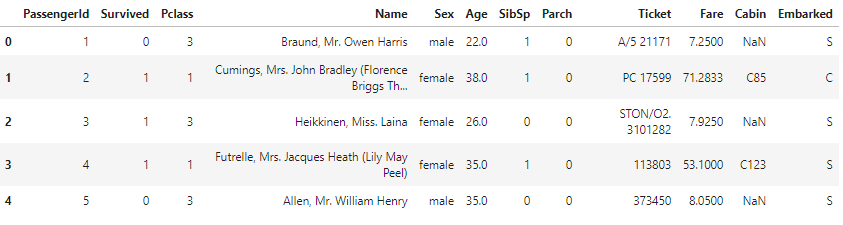
Q4

#load dataset again

titanic = pd.read\_csv("Titanic-Dataset.csv")

titanic.head(5)

4\_1

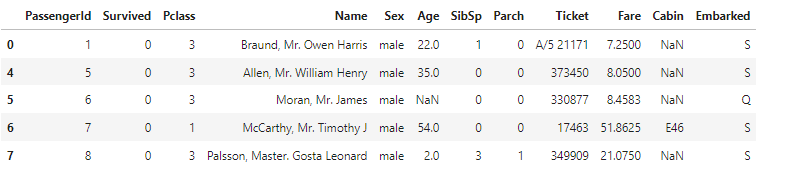


# Filter non survived male

titanic\_filter = titanic[(titanic["Survived"]==0)&(titanic["Sex"]=="male")]

titanic\_filter.head(5)

4\_2



# Calculate mean,Median,Mode,Quartile Ranges, Standard deviation of age for "Not Survived Male"

mean = titanic\_filter["Age"].mean()

print("mean age for Not Survived Male - ",mean)

median = titanic\_filter["Age"].median()

print("median age for Not Survived Male - ",median)

mode = titanic\_filter["Age"].mode()[0]

print("mode age for Not Survived Male - ",mode)

q1 = titanic\_filter["Age"].quantile(0.25)

print("25% quartile age for Not Survived Male - ",q1)

q3 = titanic\_filter["Age"].quantile(0.75)

print("75% quartile age for Not Survived Male - ",q3)

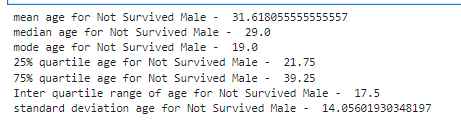
iqr = q3-q1

print("Inter quartile range of age for Not Survived Male - ",iqr)

std = titanic\_filter["Age"].std()

print("standard deviation age for Not Survived Male - ",std)

4\_3



Q5

# Correlation between Survived, Parch and Age

sns.heatmap(titanic[["Survived","Parch","Age"]].corr(),annot = True,cmap='viridis')

5\_1

