PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Ft. Road, BSK III Stage, Bengaluru – 560 085

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Course Title: Image Processing and Data Visualization Using MATLAB	
Course code: -UE19CS257B	
SRN: PES1UG19CS321	**Name: PALAK KOTHARI**

PROJECT REPORT

Problem Statement: Who Survives the titanic?

Objectives: To build a predictive model that answers the question: "What sorts of people were more likely to survive?" Using passenger data (i.e name, age, gender, socio-economic class, etc)

Description:

About the Dataset:

- Each row in the dataset represents a passenger or a crew member
- Each column has various attributes of the passenger such as passenger's name, age, sex, number of children/parents, number of siblings/spouses, port of embarkation, ticket number, fare, etc.
- There are 891 rows and 12 columns

Description of Dataset:

PASSENGER ID	Unique ID used to identify the passenger
SURVIVED	Value = 1 if passenger survived and Value = 0 if passenger did not survive

PCLASS	Ticket Class (1=1st, 2=2nd, 3=3rd)
NAME	Name of the passenger
SEX	Sex of the passenger
AGE	Age of the passenger
SIBSP	Number of siblings/ spouses of the passenger present on the Titanic
PARCH	Number of parents/children present on the Titanic
TICKET	Ticket number of the passenger
FARE	Passenger fare
CABIN	Cabin number
EMBARKED	Port of embarkation (C=Cherbourg, Q=Queenstown, S=Southampton)

Uncleaned Data:

	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

Cleaned Data:

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

New Concept Learnt (Explanation):

- Data cleaning: Checking for missing Values and Filling them with appropriate average values, and checking for duplicate values and deleting them.
- Data Visualization: Using graphs, we plot various the variables (columns) against the survival rates to try and understand the relationship between each factor.
- Outlier Filtering: How to filter the unwanted outliers, if any.
- Normalization: We normalized the data by calculating the measure of central tendencies for each variable and plotted the new data.
- Hypothesis Testing: We assumed a few hypotheses and ran tests on them to check their applicability.

Learning outcome:

We learnt how to MATLAB's vast functionalities and application in a statistical background.
Handling data and analyzing its various characteristics have never been easier before.

```
disp(countofmissingvalues);

Embarked_mode=mode(Embarked);
ds=fillmissing(ds,'constant',Embarked_mode,"DataVariables",'Embarked');
ix = ismissing(ds);
countofmissingvalues=sum(ix);
disp(countofmissingvalues);
```

```
%DATA IMPORTING
ds=readtable("titanic_train_uncleaned.csv");

variableNames = ds.Properties.VariableNames;
%ds.(variableNames{9}) = char(ds.(variableNames{9}));
ds.(variableNames{12}) = char(ds.(variableNames{12}));
Survived=ds.Survived;
Pclass=ds.Pclass;
Ticket=ds.Ticket;
Age=ds.Age;
Embarked=ds.Embarked;

%DATA CLEANING
ix = ismissing(ds);
countofmissingvalues=sum(ix);
disp(countofmissingvalues);

ds=fillmissing(ds,'next',"DataVariables",'Ticket');
%ds=fillmissing(ds,'constant',0,"DataVariables",'Ticket');
ix = ismissing(ds);
countofmissingvalues=sum(ix);
disp(countofmissingvalues);

ds=fillmissing(ds,'constant',0,"DataVariables",'Age'); %replace missing values in age with 0 to calculate mean
mean_age=mean(int8(Age));
ds.Age(ds.Age==0)=mean_age; %replace previously placed 0 to mean age value
ix = ismissing(ds);
countofmissingvalues=sum(ix);
disp(countofmissingvalues);

Embarked_mode=mode(Embarked);
ds=fillmissing(ds,'constant',Embarked_mode,"DataVariables",'Embarked');
ix = ismissing(ds);
countofmissingvalues=sum(ix);
```

We learnt how to clean, filter and visualize large Datasets that will provide a valuable input to various other fields of Computer Science such as Machine Learning.

We learnt the concept of Hypothesis Testing better and how to apply it practically.

Code:

```
%DATA VISUALISATION
figure(1);
cSurv=categorical(ds.Survived,[0,1],{'0','1'});
histogram(cSurv);
xlabel('Survived');ylabel('Count');

figure(2);
cSex=categorical(ds.Sex,[0,1],{'female','male'});
histogram(cSex,'FaceColor','green');
xlabel('Sex');ylabel('Count');
figure(3);
heatmap(ds,'Sex','Survived');

figure(4);
histogram(ds.Pclass,'FaceColor','red');
xlabel('Pclass');ylabel('Count');
figure(5);
heatmap(ds,'Survived','Pclass');

figure(6);
histogram(ds.Fare,'FaceColor','yellow');xlabel('Fare');

figure(7);
histogram(ds.Age,'FaceColor','cyan');xlabel('Age');

figure(8);
histogram(ds.SibSp,'FaceColor','magenta');xlabel('SibSp');
figure(9);
heatmap(ds,'Survived','SibSp');

figure(10);
histogram(ds.Parch,'FaceColor','None');xlabel('Parch');
figure(11);
heatmap(ds,'Survived','Parch');
```

```
% Fill outliers
[newTable,outlierIndices,thresholdLow,thresholdHigh] = filloutliers(dscopy,...
    "spline","movmean",20,"DataVariables","Age");

% Display results
clf
plot(dscopy.Age,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(newTable.Age,"Color",[0 114 189]/255,"LineWidth",1.5,...
    "DisplayName","Cleaned data")

% Plot outliers
plot(find(outlierIndices(:,6)),dscopy.Age(outlierIndices(:,6)),"x",...
    "Color",[64 64 64]/255,"DisplayName","Outliers")
title("Number of outliers: " + nnz(outlierIndices(:,6)))

% Plot filled outliers
plot(find(outlierIndices(:,6)),newTable.Age(outlierIndices(:,6)),".",...
    "MarkerSize",12,"Color",[217 83 25]/255,"DisplayName","Filled outliers")

% Plot outlier thresholds
plot([(1:numel(dscopy.Age))'; missing; (1:numel(dscopy.Age))'],...
    [thresholdHigh.Age(:); missing; thresholdLow.Age(:)],...
    "Color",[145 145 145]/255,"DisplayName","Outlier thresholds")

hold off
legend
ylabel("Age")
clear outlierIndices thresholdLow thresholdHigh
```

```
% Fill outliers
[newTable2,outlierIndices2,thresholdLow2,thresholdHigh2] = filloutliers(newTable,...
    "spline","movmean",20,"DataVariables","Fare");

% Display results
clf
plot(newTable.Fare,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(newTable2.Fare,"Color",[0 114 189]/255,"LineWidth",1.5,...
    "DisplayName","Cleaned data")

% Plot outliers
plot(find(outlierIndices2(:,10)),newTable.Fare(outlierIndices2(:,10)),"x",...
    "Color",[64 64 64]/255,"DisplayName","Outliers")
title("Number of outliers: " + nnz(outlierIndices2(:,10)))

% Plot filled outliers
plot(find(outlierIndices2(:,10)),newTable2.Fare(outlierIndices2(:,10)),".",...
    "MarkerSize",12,"Color",[217 83 25]/255,"DisplayName","Filled outliers")

% Plot outlier thresholds
plot([(1:numel(newTable.Fare))'; missing; (1:numel(newTable.Fare))'],...
    [thresholdHigh2.Fare(:); missing; thresholdLow2.Fare(:)],...
    "Color",[145 145 145]/255,"DisplayName","Outlier thresholds")

hold off
legend
ylabel("Fare")
clear outlierIndices2 thresholdLow2 thresholdHigh2
```

Outlier Handling:

```
% Fill outliers
[newTable3,outlierIndices3,thresholdLow3,thresholdHigh3] = filloutliers(newTable2,...
    "spline","movmean",20,"DataVariables","SibSp");

% Display results
clf
plot(newTable2.SibSp,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(newTable3.SibSp,"Color",[0 114 189]/255,"LineWidth",1.5,...
    "DisplayName","Cleaned data")

% Plot outliers
plot(find(outlierIndices3(:,7)),newTable2.SibSp(outlierIndices3(:,7)),"x",...
    "Color",[64 64 64]/255,"DisplayName","Outliers")
title("Number of outliers: " + nnz(outlierIndices3(:,7)))

% Plot filled outliers
plot(find(outlierIndices3(:,7)),newTable3.SibSp(outlierIndices3(:,7)),".",...
    "MarkerSize",12,"Color",[217 83 25]/255,"DisplayName","Filled outliers")

% Plot outlier thresholds
plot([(1:numel(newTable2.SibSp))'; missing; (1:numel(newTable2.SibSp))'],...
    [thresholdHigh3.SibSp(:); missing; thresholdLow3.SibSp(:)],...
    "Color",[145 145 145]/255,"DisplayName","Outlier thresholds")

hold off
legend
ylabel("SibSp")
clear outlierIndices3 thresholdLow3 thresholdHigh3
```

```
% Fill outliers
[newTable4,outlierIndices4,thresholdLow4,thresholdHigh4] = filloutliers(newTable3,...
    "spline","movmean",20,"DataVariables","Parch");

% Display results
clf
plot(newTable3.Parch,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on
plot(newTable4.Parch,"Color",[0 114 189]/255,"LineWidth",1.5,...
    "DisplayName","Cleaned data")

% Plot outliers
plot(find(outlierIndices4(:,8)),newTable3.Parch(outlierIndices4(:,8)),"x",...
    "Color",[64 64 64]/255,"DisplayName","Outliers")
title("Number of outliers: " + nnz(outlierIndices4(:,8)))

% Plot filled outliers
plot(find(outlierIndices4(:,8)),newTable4.Parch(outlierIndices4(:,8)),".",...
    "MarkerSize",12,"Color",[217 83 25]/255,"DisplayName","Filled outliers")

% Plot outlier thresholds
plot([(1:numel(newTable3.Parch))'; missing; (1:numel(newTable3.Parch))'],...
    [thresholdHigh4.Parch(:); missing; thresholdLow4.Parch(:)],...
    "Color",[145 145 145]/255,"DisplayName","Outlier thresholds")

hold off
legend
ylabel("Parch")
clear outlierIndices4 thresholdLow4 thresholdHigh4
```

```
%NORMALISATION
disp("Mean of all the columns of the dataset");
d=["Mean of Survived:",mean(nds.Survived)];
disp(d);
d=["Mean of Pclass:",mean(nds.Pclass)];
disp(d);
d=["Mean of Sex:",mean(nds.Sex)];
disp(d);
d=["Mean of Age:",mean(nds.Age)];
disp(d);
d=["Mean of SibSp:",mean(nds.SibSp)];
disp(d);
d=["Mean of Parch:",mean(nds.Parch)];
disp(d);
d=["Mean of Fare:",mean(nds.Fare)];
disp(d);

disp("variance of all the columns of the dataset");
d=["Variance of Survived:",var(nds.Survived)];
disp(d);
d=["Variance of Pclass:",var(nds.Pclass)];
disp(d);
d=["Variance of Sex:",var(nds.Sex)];
disp(d);
d=["Variance of Age:",var(nds.Age)];
disp(d);
d=["Variance of SibSp:",var(nds.SibSp)];
disp(d);
d=["Variance of Parch:",var(nds.Parch)];
disp(d);
d=["Variance of Fare:",var(nds.Fare)];
disp(d);
```

```
data=nds(:,[1:6 8:9 11 7 10]); %rearrange to normalise
data{:, 10:end} = normc(data{:, 10:end}); %fare and age ads the last two cols

%plot after normalisation
figure(14);
histfit(data.Fare);xlabel('Fare');
figure(15);
histfit(data.Age);xlabel('Age');
nds=data;

disp("After normalisation:")
 disp("Mean of all the columns of the dataset");
d=["Mean of Survived:",mean(nds.Survived)];
disp(d);
d=["Mean of Pclass:",mean(nds.Pclass)];
disp(d);
d=["Mean of Sex:",mean(nds.Sex)];
disp(d);
d=["Mean of Age:",mean(nds.Age)];
disp(d);
d=["Mean of SibSp:",mean(nds.SibSp)];
disp(d);
d=["Mean of Parch:",mean(nds.Parch)];
disp(d);
d=["Mean of Fare:",mean(nds.Fare)];
disp(d);
```

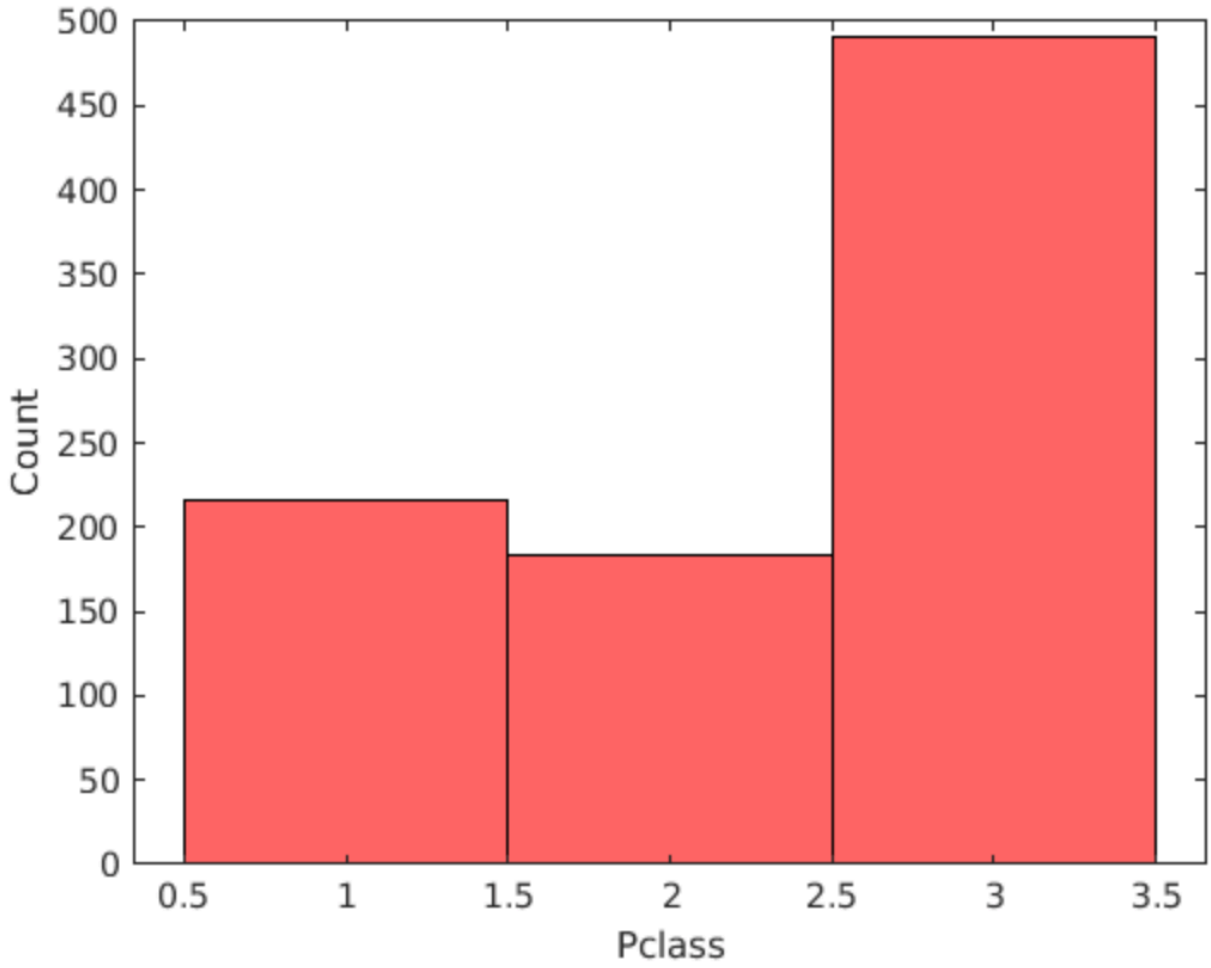
```
 disp("variance of all the columns of the dataset");
 d=["Variance of Survived:",var(nds.Survived)];
 disp(d);
 d=["Variance of Pclass:",var(nds.Pclass)];
 disp(d);
 d=["Variance of Sex:",var(nds.Sex)];
 disp(d);
 d=["Variance of Age:",var(nds.Age)];
 disp(d);
 d=["Variance of SibSp:",var(nds.SibSp)];
 disp(d);
 d=["Variance of Parch:",var(nds.Parch)];
 disp(d);
 d=["Variance of Fare:",var(nds.Fare)];
 disp(d);
```

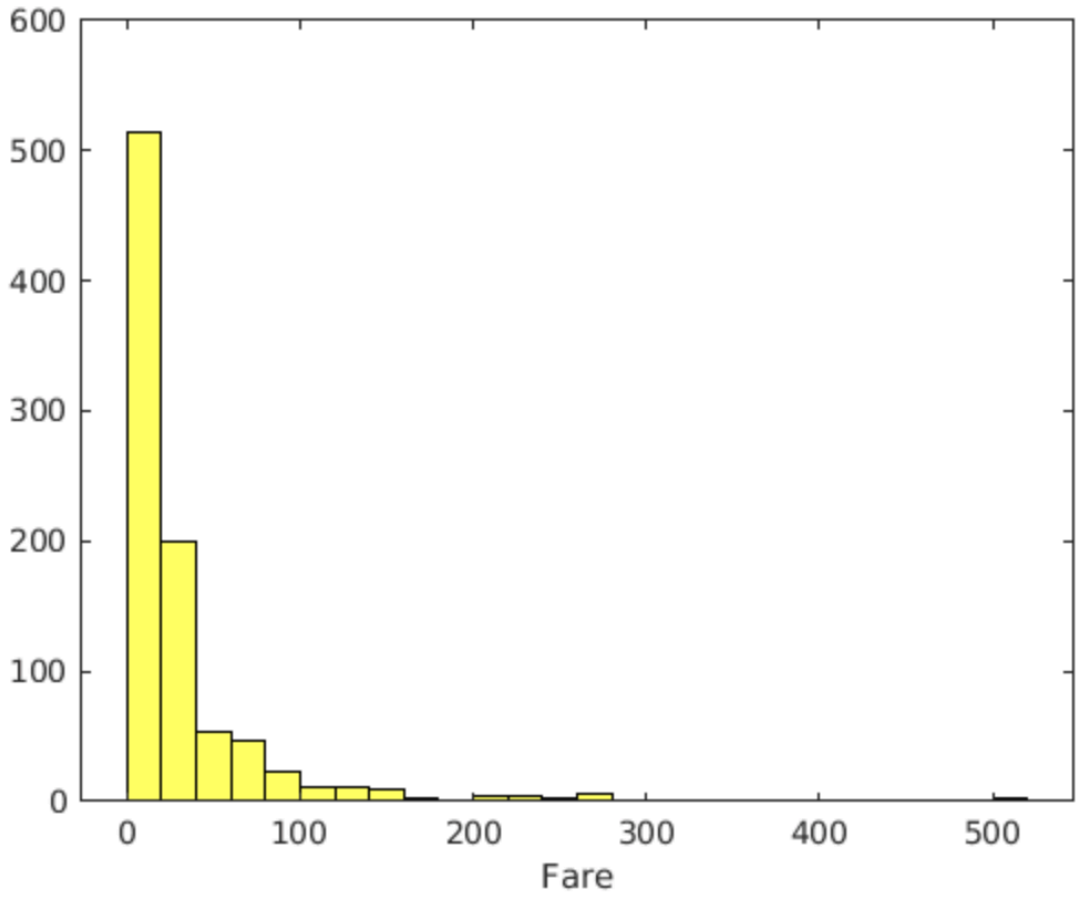
```
%HYPOTHESIS TESTING

disp("Running a Two-sample Kolmogorov-Smirnov Hypothesis Test: with alpha value 5%")
disp("HYPOTHESIS: The proportion of females onboard who survived the sinking of the Titanic was higher " + ...
    "than the proportion of males onboard who survived the sinking of the Titanic.");
disp("NULL HYPOTHESIS: There is no relationship between the sex and the survived or The proportion of " + ...
    "female survivors is equal to the proportion of male survivors");
disp("ALTERNATE HYPOTHESIS: There is a relationship between the sex and the survived or The proportion " + ...
    "of female survivors is not equal to the proportion of male survivors");
[h,p]=kstest2(data.Survived,data.Sex);

if (p<0.05)
    disp("The null hypothesis is rejected!!! The sex and survived class are related!")
else
    disp("The null hypothesis is not rejected!!! The sex and survived class are not related!");
end

%CORRELATION : TO FURTHER SUPPORT OUR HYPOTHESIS

corr=[data.Survived,data.Sex,data.Pclass,data.SibSp,data.Parch,data.Age,data.Fare,];
corrplot(corr,'varNames',{'Survived','Sex','Pclass','SibSp','Parch','Age','Fare'});
```

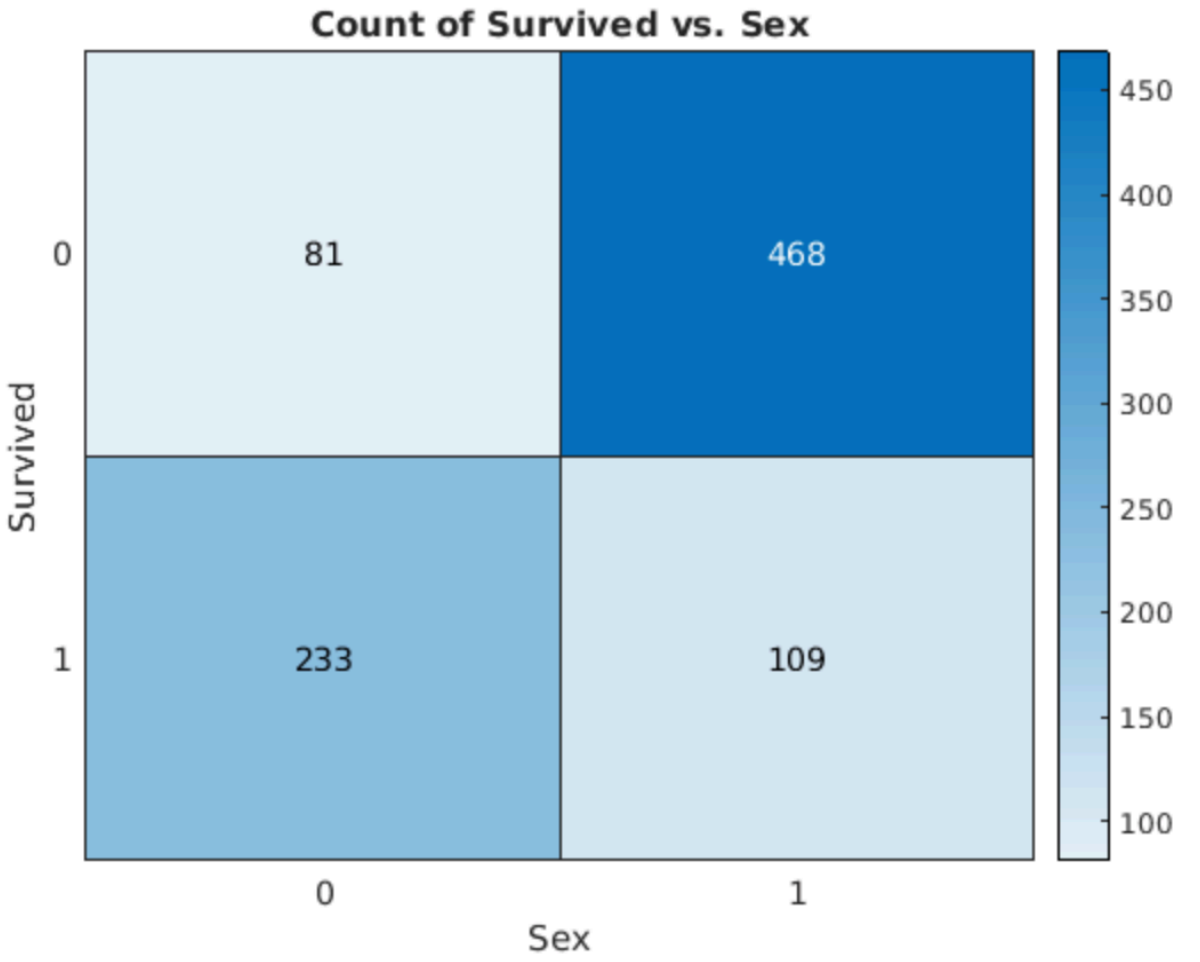
Output Screenshots:

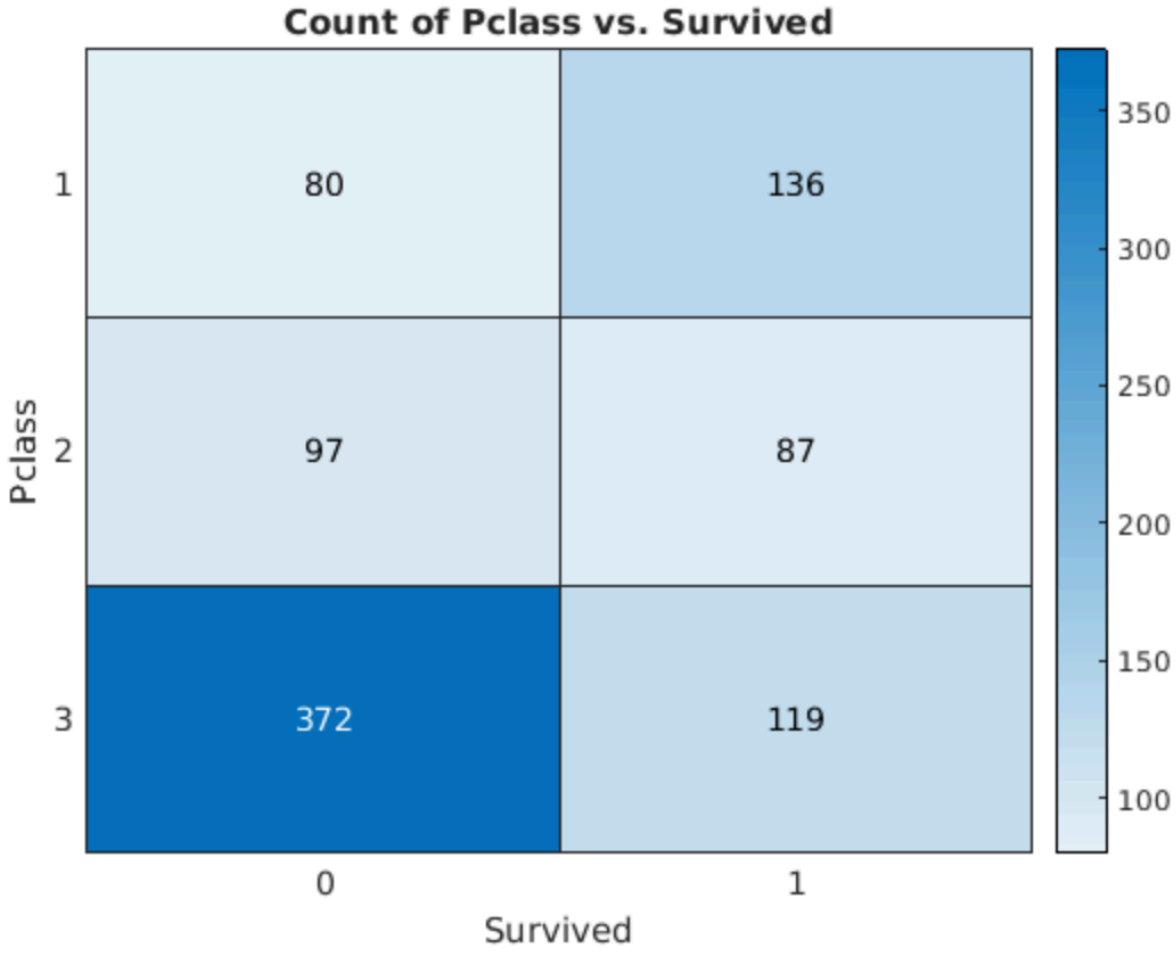
Missing Values and Duplicate Values Handling;

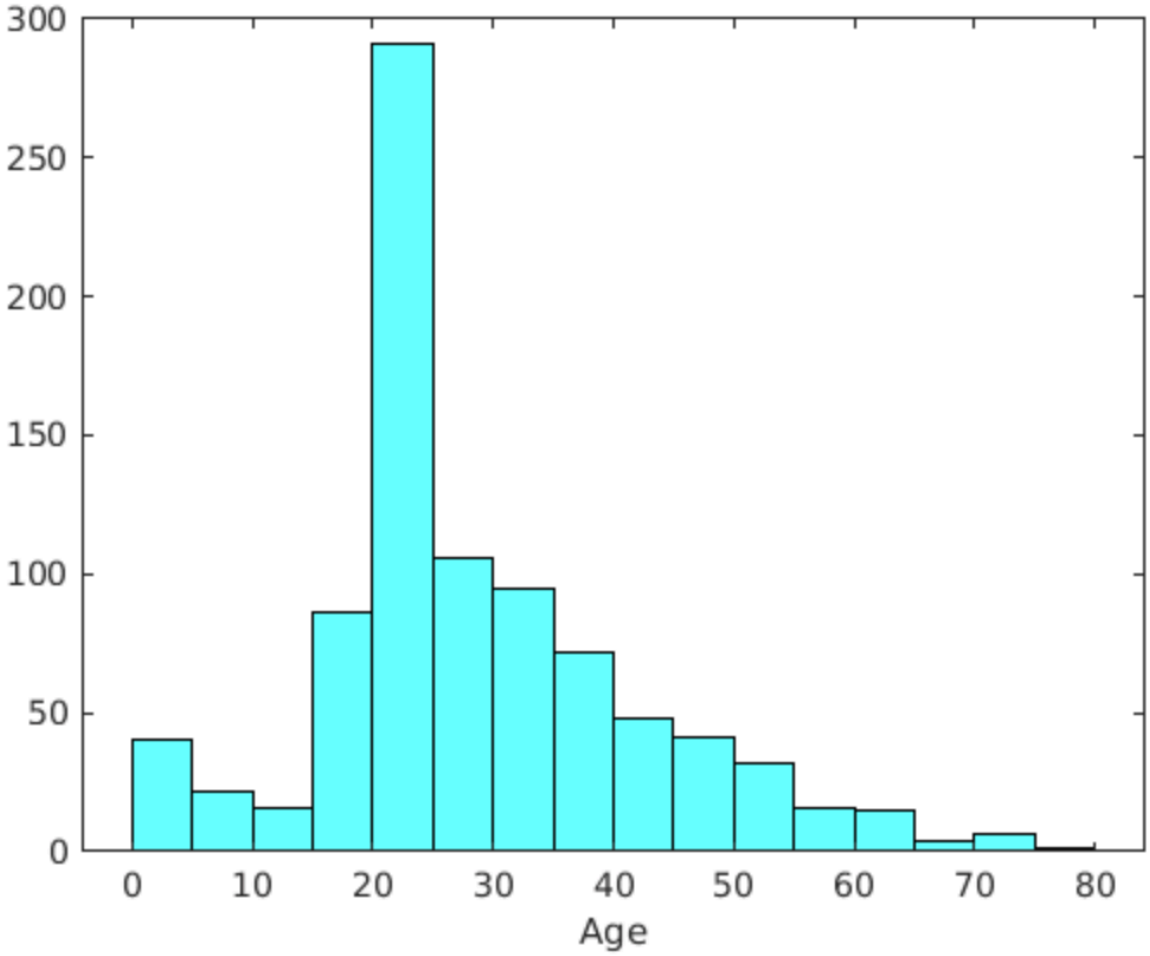
```
     0     0     0     0     0   177     0     0   230     0   687     2

     0     0     0     0     0   177     0     0     0     0   687     2

     0     0     0     0     0     0     0     0     0     0   687     2

     0     0     0     0     0     0     0     0     0     0   687     0

     0     0     0     0     0     0     0     0     0     0     0

NO DUPLICATES FOUND
```


500
450
400
350
300
250
200
150
100
50
0
Count
0.5
1
1.5
2
2.5
3
3.5
Pclass


600
500
400
300
200
100
0
0
100
200
300
400
500
Fare

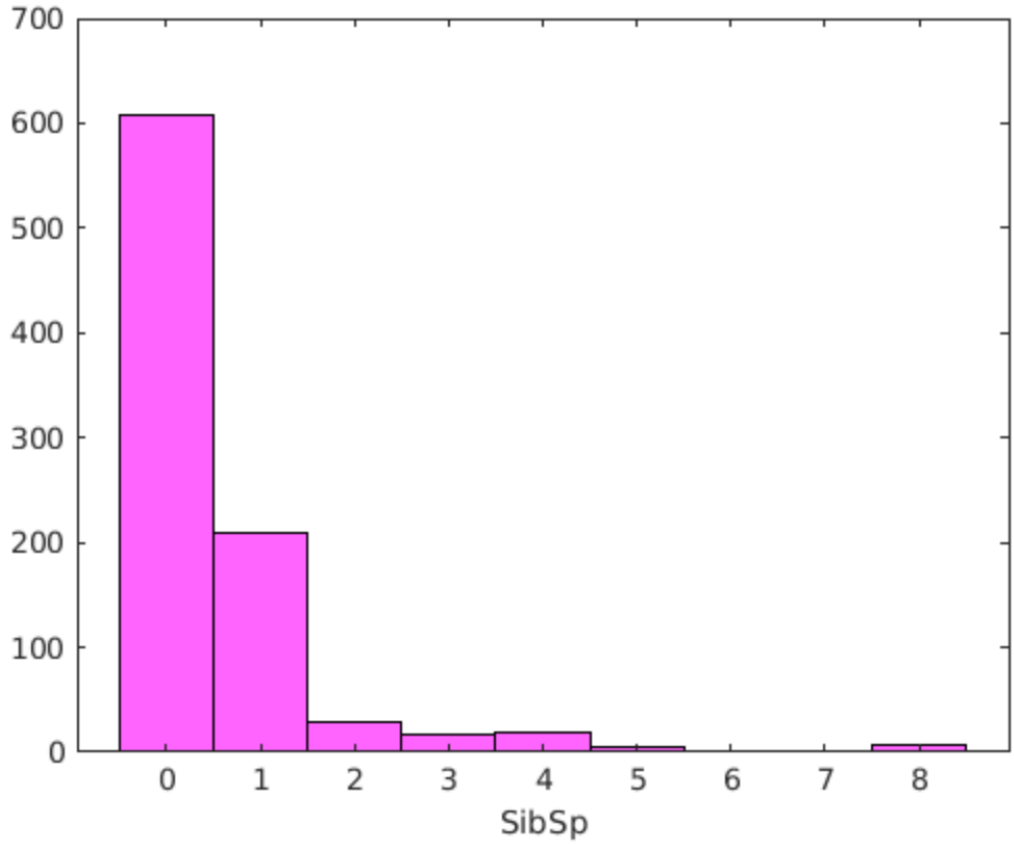

Count of Survived vs. Sex
Survived
0
1
81
468
233
109
0
1
Sex
450
400
350
300
250
200
150
100

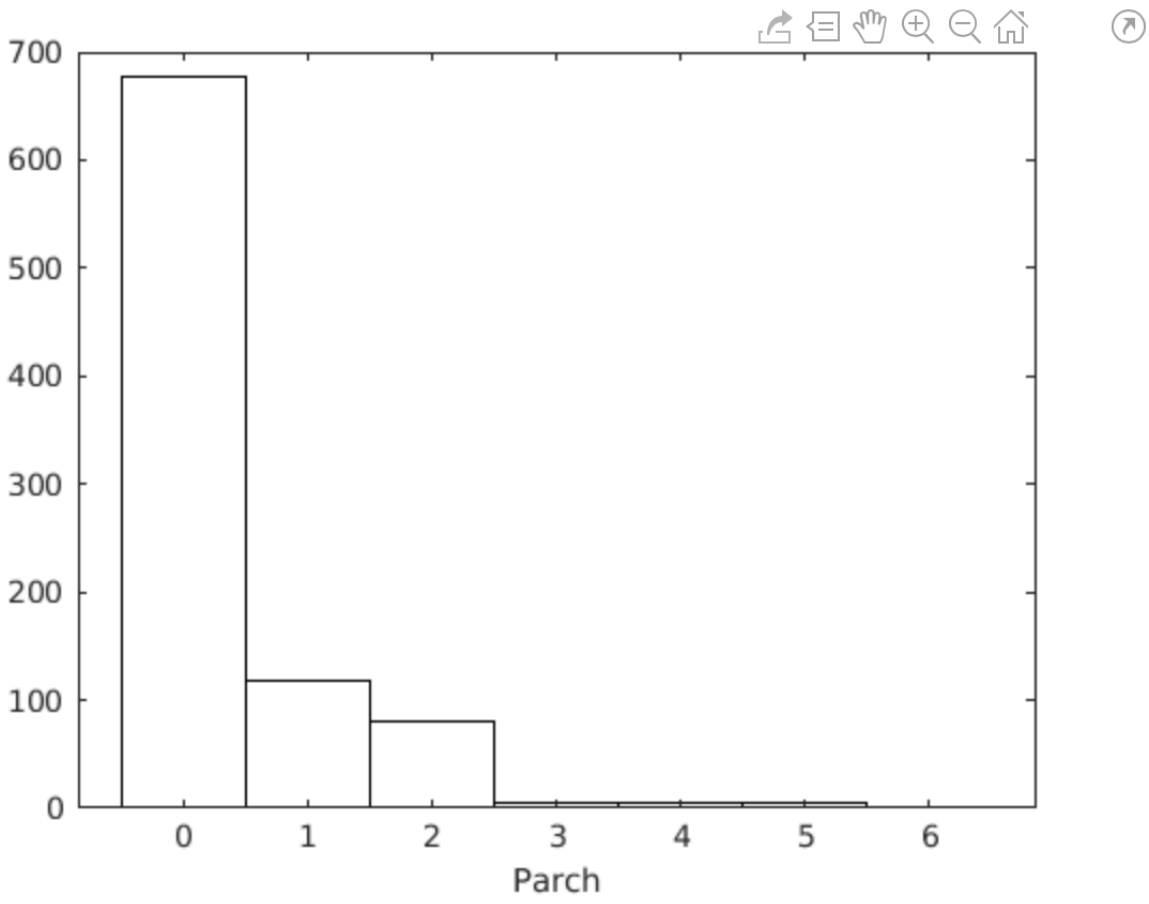

Count of Pclass vs. Survived
Pclass
1
2
3
80
136
97
87
372
119
0
1
Survived
350
300
250
200
150
100

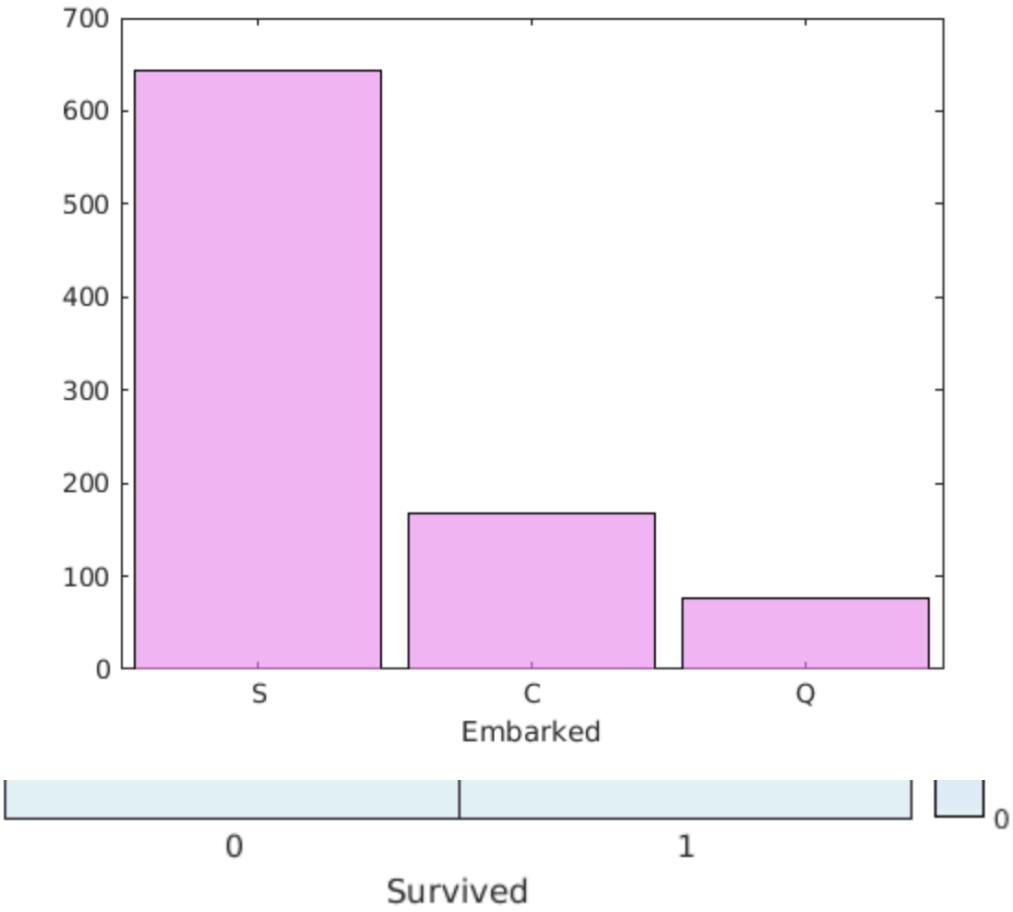


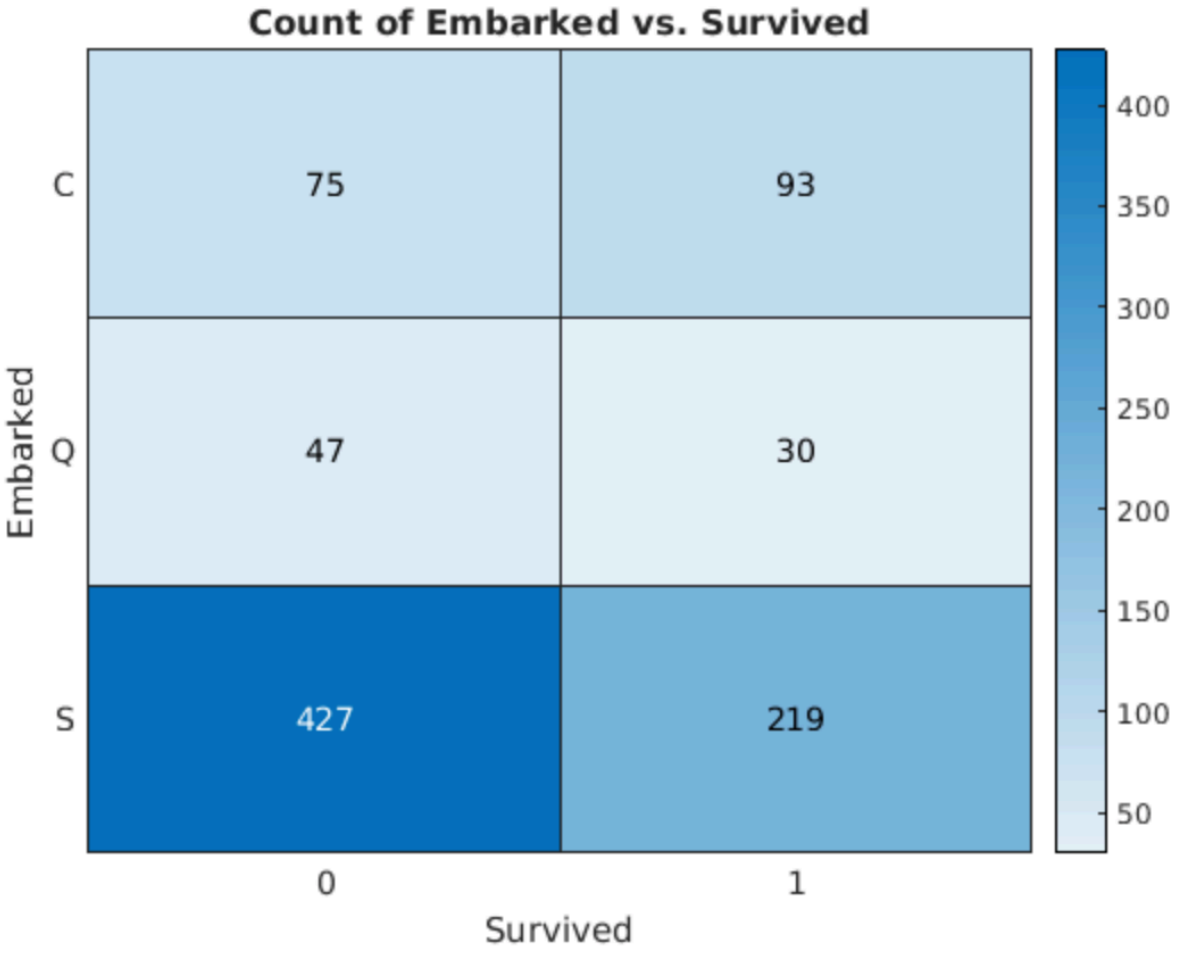
Count of SibSp vs. Survived

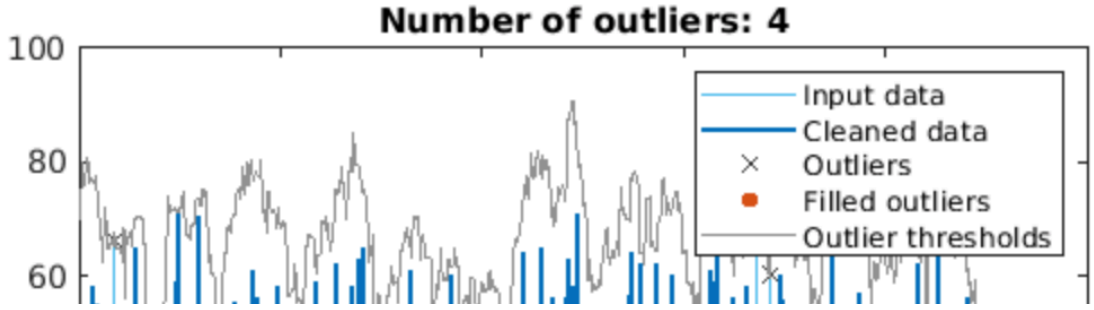
SibSp	Survived = 0	Survived = 1
0	398	210
1	97	112
2	15	13
3	12	4
4	15	3
5	5	0
8	7	0

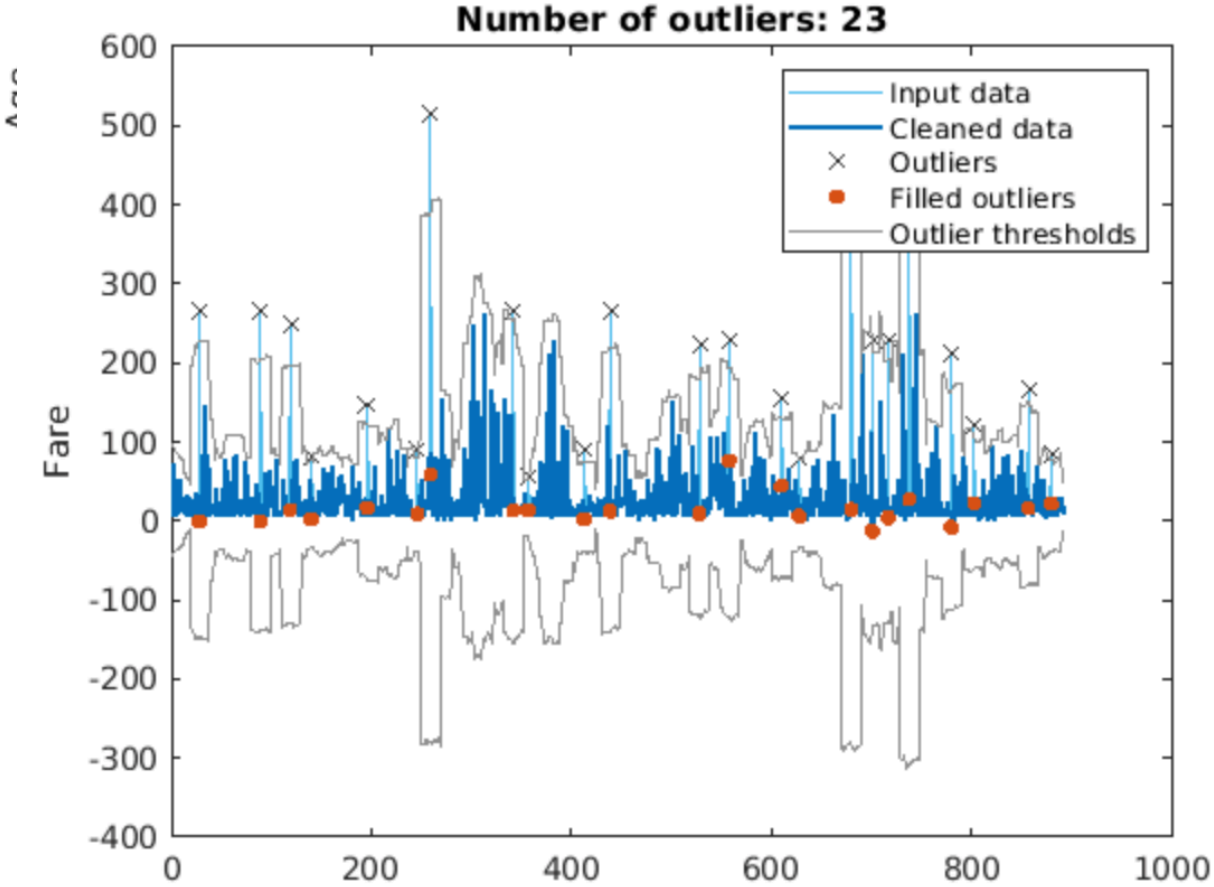

700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
SibSp

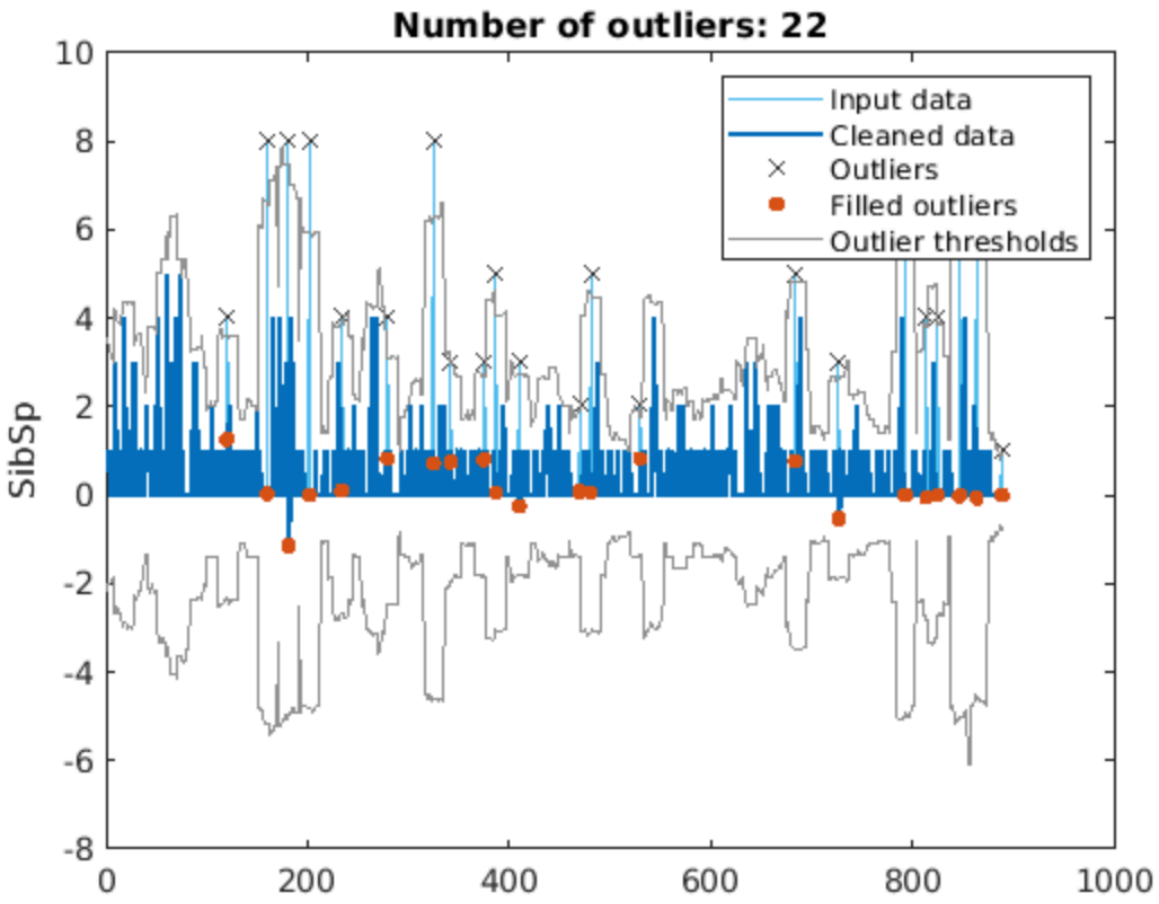

700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
Parch

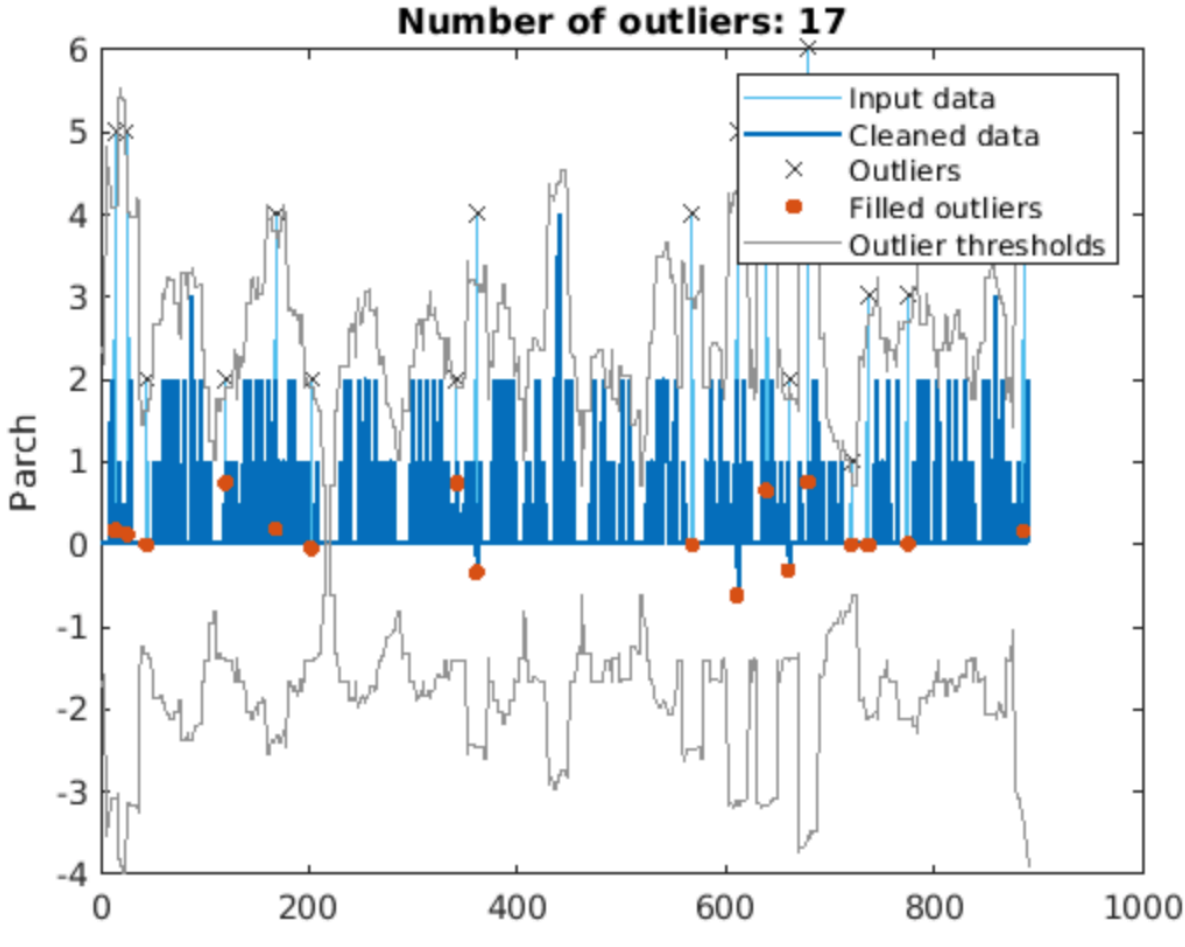

700
600
500
400
300
200
100
0
S
C
Q
Embarked
0
1
Survived
0


Count of Embarked vs. Survived
C
Q
S
Embarked
75
93
47
30
427
219
0
1
Survived
400
350
300
250
200
150
100
50


Number of outliers: 4
100
80
60
Input data
Cleaned data
Outliers
Filled outliers
Outlier thresholds

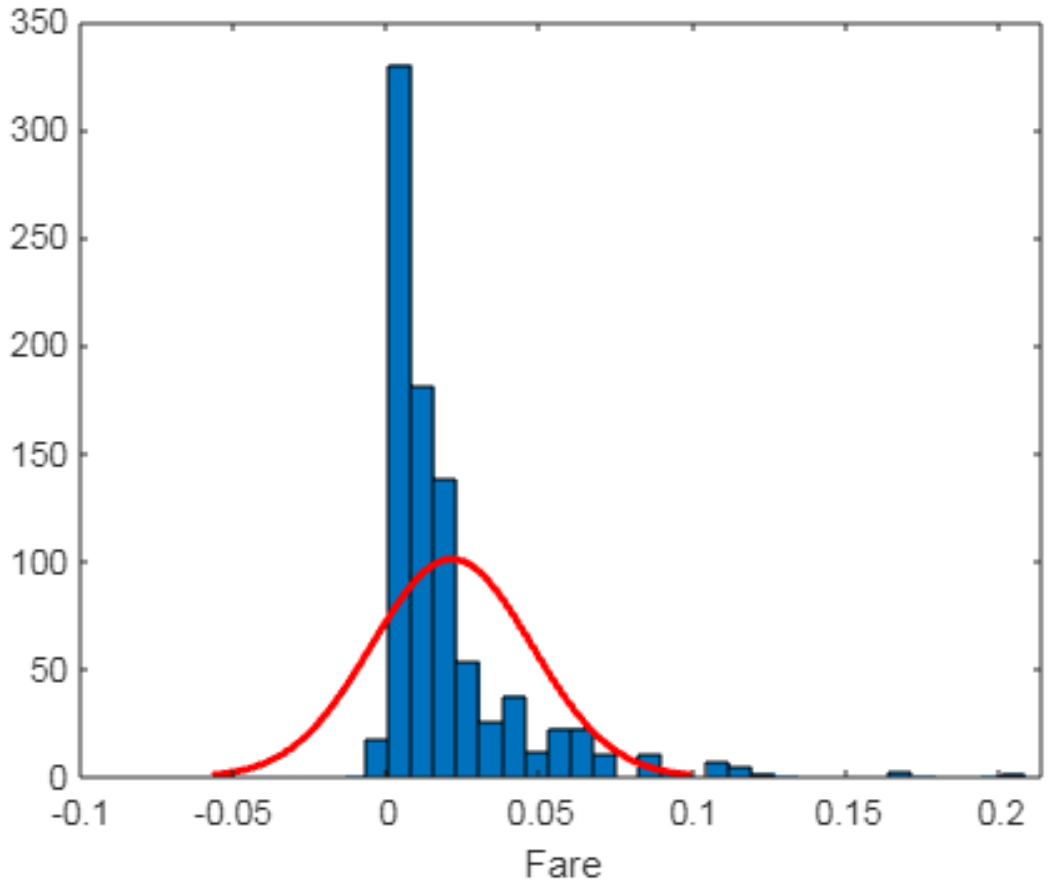

Number of outliers: 23
Fare
600
500
400
300
200
100
0
-100
-200
-300
-400
0
200
400
600
800
1000
Input data
Cleaned data
Outliers
Filled outliers
Outlier thresholds

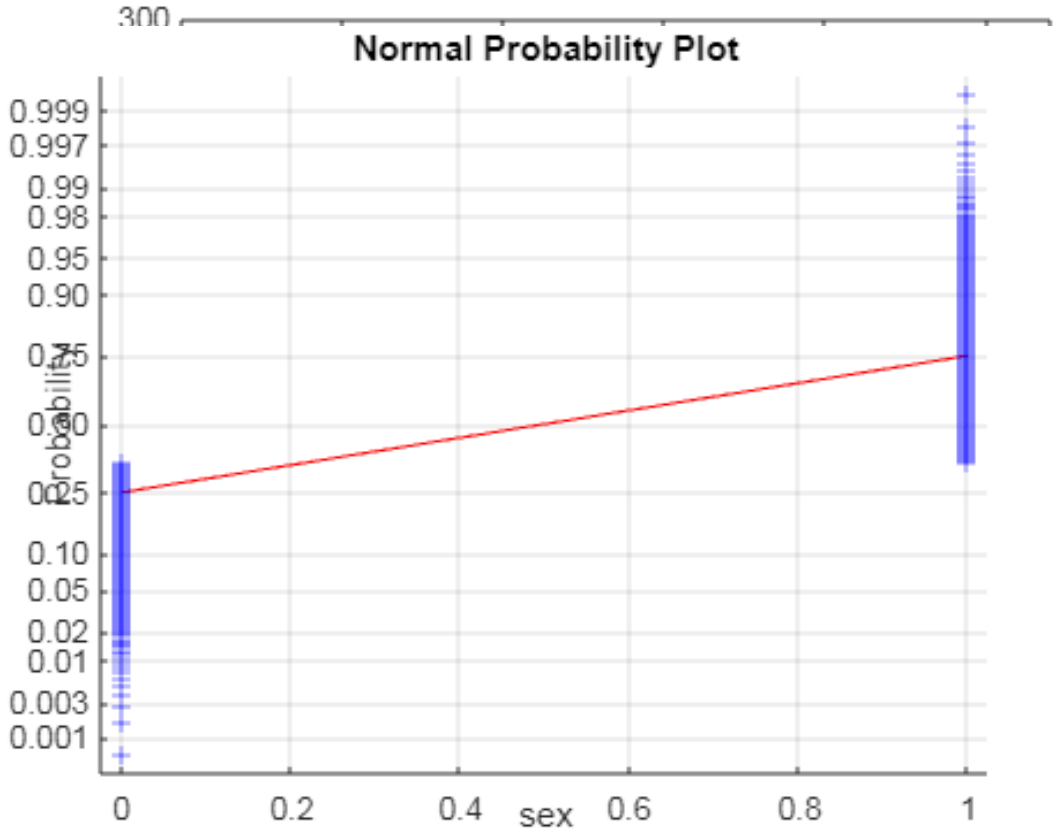

Number of outliers: 22
Input data
Cleaned data
Outliers
Filled outliers
Outlier thresholds
SibSp
10
8
6
4
2
0
-2
-4
-6
-8
0
200
400
600
800
1000

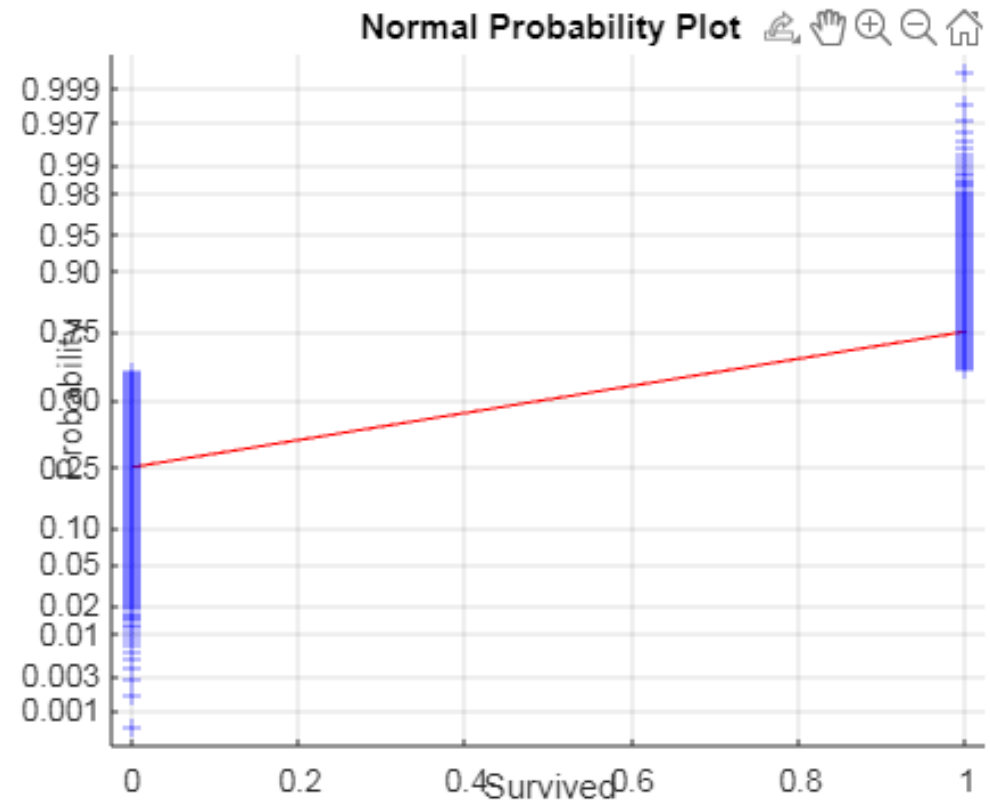

Number of outliers: 17
Input data
Cleaned data
Outliers
Filled outliers
Outlier thresholds
Parch
6
5
4
3
2
1
0
-1
-2
-3
-4
0
200
400
600
800
1000

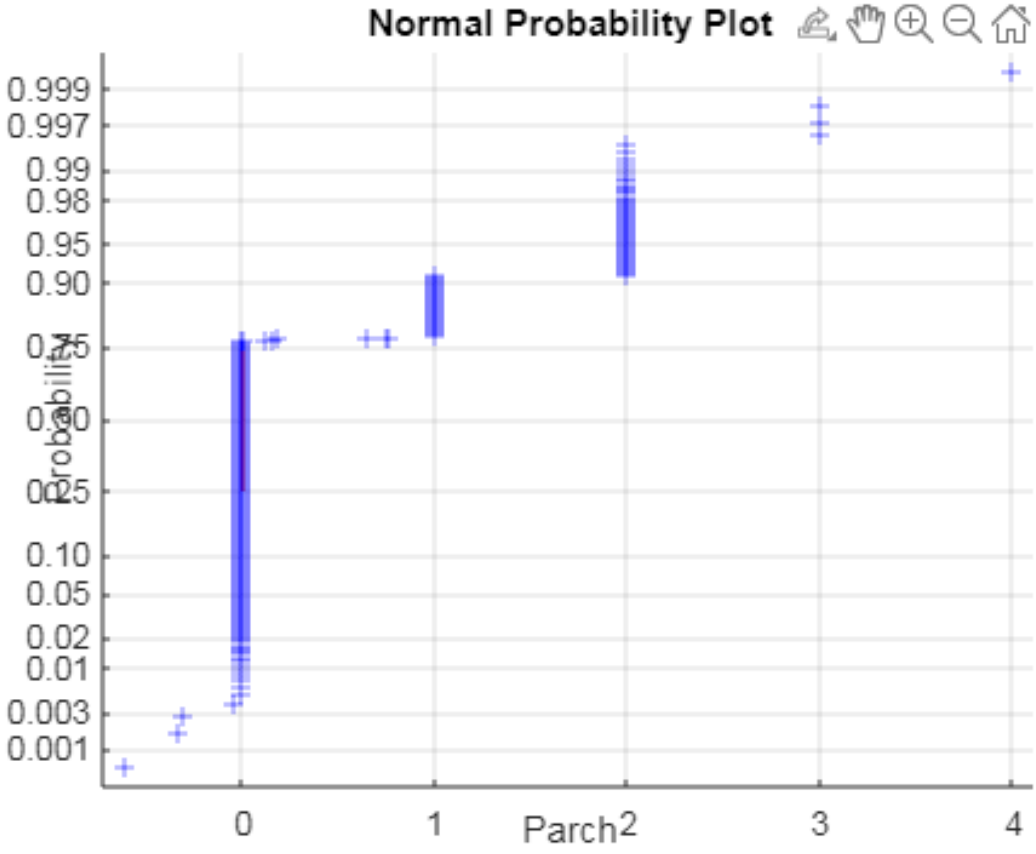
```
Mean of all the columns of the dataset

    "Mean of Survived:"    "0.38384"

    "Mean of Pclass:"    "2.3086"

    "Mean of Sex:"    "0.64759"

    "Mean of Age:"    "28.423"

    "Mean of SibSp:"    "0.40657"

    "Mean of Parch:"    "0.31679"

    "Mean of Fare:"    "26.9845"

variance of all the columns of the dataset

    "Variance of Survived:"    "0.23677"

    "Variance of Pclass:"    "0.69902"

    "Variance of Sex:"    "0.22847"

    "Variance of Age:"    "169.1763"

    "Variance of SibSp:"    "0.60417"

    "Variance of Parch:"    "0.41955"

    "Variance of Fare:"    "1110.3932"
```

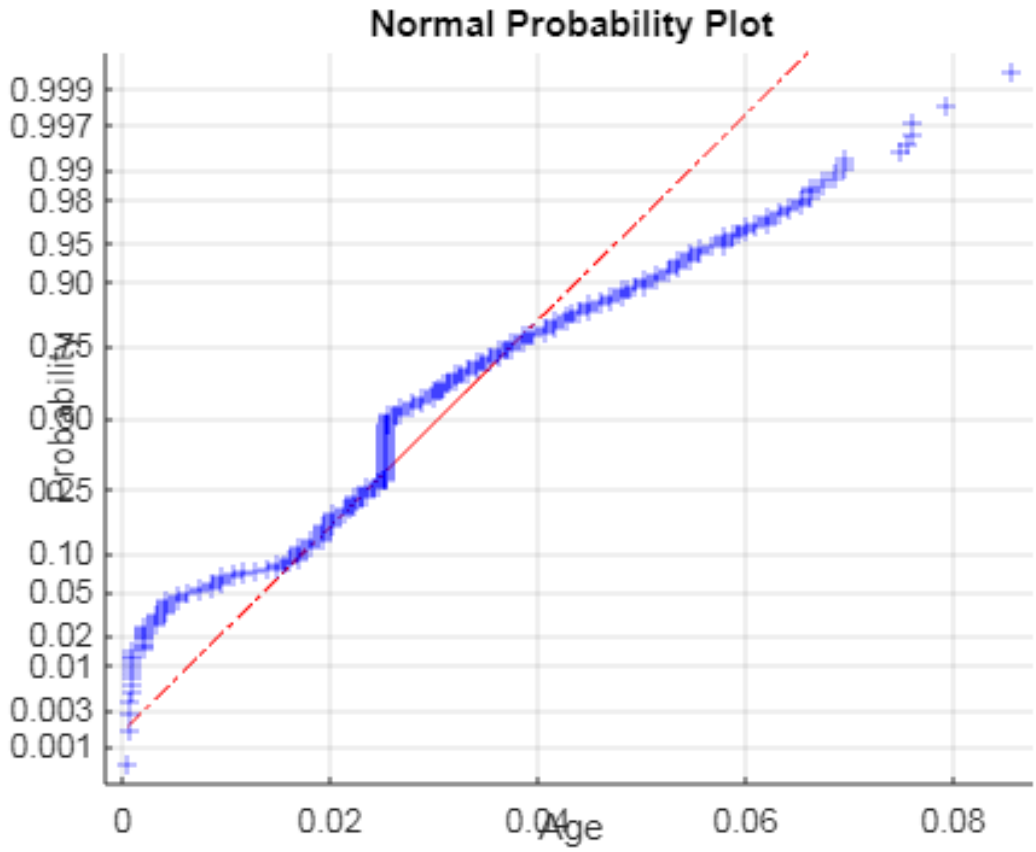
```
After normalisation:
Mean of all the columns of the dataset
    "Mean of Survived:"    "0.38384"

    "Mean of Pclass:"    "2.3086"

    "Mean of Sex:"    "0.64759"

    "Mean of Age:"    "0.030466"

    "Mean of SibSp:"    "0.40657"

    "Mean of Parch:"    "0.31679"

    "Mean of Fare:"    "0.02109"



variance of all the columns of the dataset
    "Variance of Survived:"    "0.23677"

    "Variance of Pclass:"    "0.69902"

    "Variance of Sex:"    "0.22847"

    "Variance of Age:"    "0.00019437"

    "Variance of SibSp:"    "0.60417"

    "Variance of Parch:"    "0.41955"

    "Variance of Fare:"    "0.00067829"
```

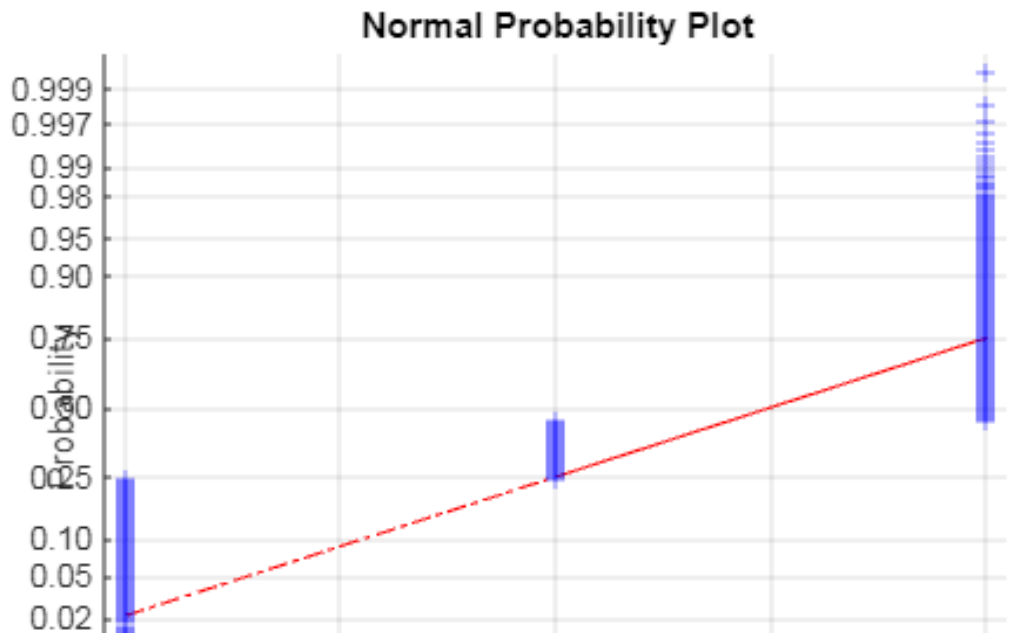

350
300
250
200
150
100
50
0
-0.1
-0.05
0
0.05
0.1
0.15
0.2
Fare


300
Normal Probability Plot
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
Probability
0
0.2
0.4
sex
0.6
0.8
1


Normal Probability Plot
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
Probability
0
0.2
0.4
Survived
0.6
0.8
1


Normal Probability Plot
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
Probability
0
1
Parch
2
3
4


Normal Probability Plot
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
Probability
0
0.02
0.04
0.06
0.08
Age



Running a Two-sample Kolmogorov-Smirnov Hypothesis Test: with alpha value 5%

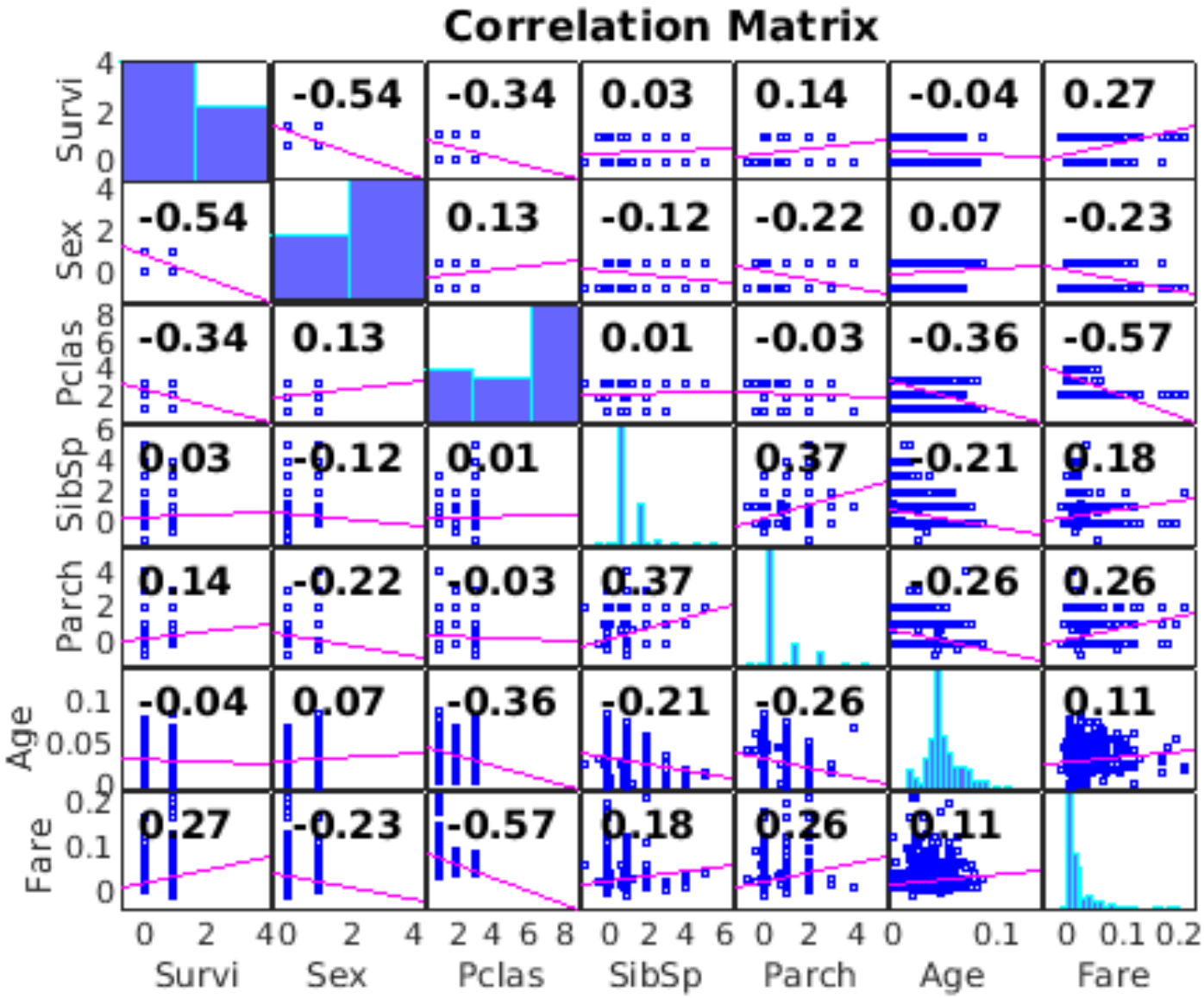
HYPOTHESIS: The proportion of females onboard who survived the sinking of the Titanic was higher than the proportion of males onboard who survived the sinking of the Titanic.

NULL HYPOTHESIS: There is no relationship between the sex and the survived or The proportion of female survivors is equal to the proportion of male survivors

ALTERNATE HYPOTHESIS: There is a relationship between the sex and the survived or The proportion of female survivors is not equal to the proportion of male survivors

The null hypothesis is rejected!!! The sex and survived class are related!

Running a Two-sample Kolmogorov-Smirnov Hypothesis Test: with alpha value 5%
HYPOTHESIS: The proportion of females onboard who survived the sinking of the Titanic was higher than the proportion of males onboard who survived the sinking of the Titanic.
NULL HYPOTHESIS: There is no relationship between the sex and the survived or The proportion of female survivors is equal to the proportion of male survivors
ALTERNATE HYPOTHESIS: There is a relationship between the sex and the survived or The proportion of female survivors is not equal to the proportion of male survivors
The null hypothesis is rejected!!! The sex and survived class are related!



Name and signature of the faculty: Prof. Revathi