**Name: Parth Das**

**Sap: 60004220185**

**Roll No: C-111**

**Subject: Business Analytics**

**Exp 1:** Import appropriate dataset of your choice. Filter/clean the data by removing missing values /outliers. Do appropriate transformations to add meaningful columns. Export it to a pdf file.

**Theory**:- SAS Code: This SAS script performs the following steps:

1. Imports the Titanic dataset from a public URL using PROC IMPORT.

2. Cleans the data by: - Removing rows with any missing values using CMISS. - Handling outliers in the Age and Fare columns based on the Interquartile Range (IQR) method with PROC UNIVARIATE.

3. Transforms the data by adding: - FamilySize (combining SibSp and Parch plus the passenger). - IsMinor to indicate if the passenger is under 18.

4. Exports a summary report to a PDF file using ODS PDF, including descriptive statistics (PROC MEANS) and a sample of the cleaned dataset (PROC PRINT).

**Code:-**

/\* Step 1: Import Dataset \*/

filename titanic url

'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv';

proc import datafile=titanic

out=titanic\_data

dbms=csv

replace;

getnames=yes;

run;

/\* Step 2: Data Cleaning \*/

/\* Removing missing values \*/

data titanic\_clean;

set titanic\_data;

if cmiss(of \_all\_) = 0; /\* Removes rows with any missing values \*/

run;

/\* Handling outliers using IQR method \*/

proc univariate data=titanic\_clean noprint;

var Age Fare;

output out=iqr\_data pctlpre=P\_ pctlpts=25,75;

run;

data titanic\_clean;

merge titanic\_clean iqr\_data;

by \_all\_;

IQR\_Age = P\_75 - P\_25;

lower\_bound\_age = P\_25 - 1.5 \* IQR\_Age;

upper\_bound\_age = P\_75 + 1.5 \* IQR\_Age;

if Age >= lower\_bound\_age and Age <= upper\_bound\_age;

IQR\_Fare = P\_75 - P\_25;

lower\_bound\_fare = P\_25 - 1.5 \* IQR\_Fare;

upper\_bound\_fare = P\_75 + 1.5 \* IQR\_Fare;

if Fare >= lower\_bound\_fare and Fare <= upper\_bound\_fare;

run;

/\* Step 3: Data Transformation \*/

data titanic\_transformed;

set titanic\_clean;

FamilySize = SibSp + Parch + 1;

if Age < 18 then IsMinor = 1;

else IsMinor = 0;

run;

/\* Step 4: Export to PDF \*/

ods PDF file="cleaned\_titanic\_data\_summary.pdf";

proc means data=titanic\_transformed n mean std min max;

var Age Fare FamilySize;

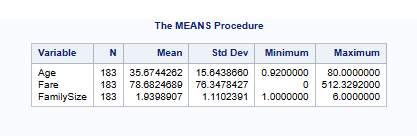
run;

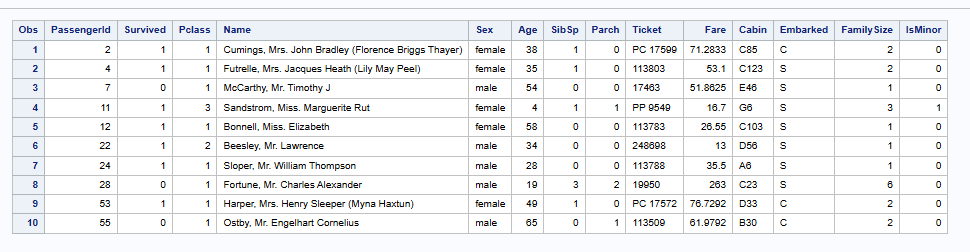
proc print data=titanic\_transformed (obs=10);

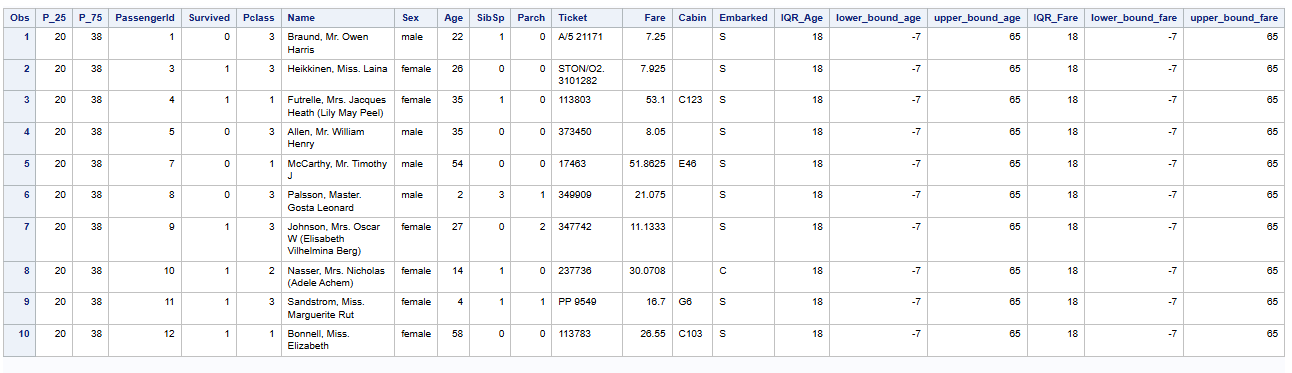
run;

ods PDF close;

**Results:-**







**Conclusion:-**

In this experiment, we successfully imported the Titanic dataset, performed data cleaning by removing missing values and handling outliers using the Interquartile Range (IQR) method. We then transformed the data by adding meaningful columns such as **FamilySize** and **IsMinor** to enhance our analysis. Finally, we generated a summary report containing descriptive statistics and a sample of the cleaned dataset, exporting it to a **PDF file** for documentation and further insights.

This approach demonstrates the importance of **data preprocessing** in real-world scenarios, ensuring data quality before applying analytical models or machine learning algorithms.