Nume: Ariciu Toma Grupă: 312CAb

Proiect PCLP3 - Titanic - Partea I

Descriere:

Cerinta 1

- Am importat csv-ul cu ajutorul bibliotecii Pandas, iar apoi folosind functii din cadrul acesteia, am determinat:
- Numarul de coloane: 11
- Tipurile datelor din coloane: Survived int64 Pclass int64 Name object Sex object Age float64 SibSp int64 Parch int64 Ticket object Fare float64 Cabin object Embarked object
- Numarul de valori lipsa din fiecare coloana: Survived 0 Pclass 0 Name 0 Sex 0 Age 177 SibSp 0 Parch 0 Ticket 0 Fare 0 Cabin 687 Embarked 2
- Numarul de linii: 891
- Numarul de linii duplicate: 0

Cerinta 2

- Am folosit functia value_counts() pentru a calcula numarul aparitiilor din fiecare coloana ceruta - Survived, Pclass, Sex. Apoi, le-am pus in pie-charturi folosind biblioteca matplotlib.

Cerinta 3

- Am luat pe rand fiecare coloana cu valori numerice, am eliminat valorile lipsa cu ajutorul functiei dropna(), iar apoi am generat histograma corespunzatoare:

Cerinta 4

- Am verificat pentru fiecare coloana daca are valori lipsa si, daca da, cate sunt si ce procentaj reprezinta.
- Age: Numarul total este 177 Procentajul este 19.865319865319865%
- Cabin: Numarul total este 687 Procentajul este 77.10437710437711%
- Embarked: Numarul total este 2 Procentajul este 0.2244668911335783%
- Am facut aceeasi verificare luand in parte cele doua clase date de coloana Survived: cei care au supravietuit si cei care nu.

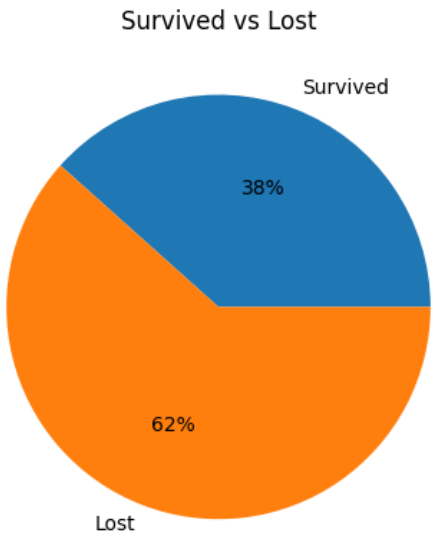

Figure 1: plot_survivors

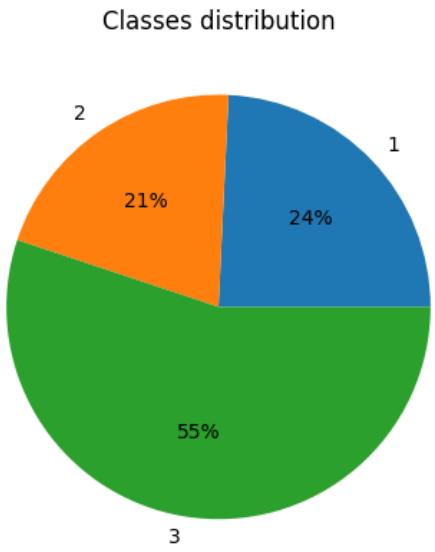

Figure 2: plot_classes

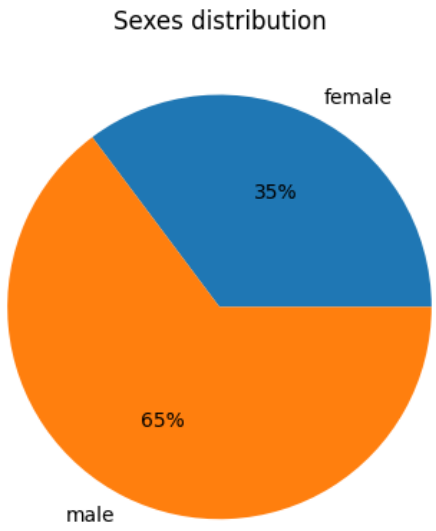

Figure 3: plot_sexes

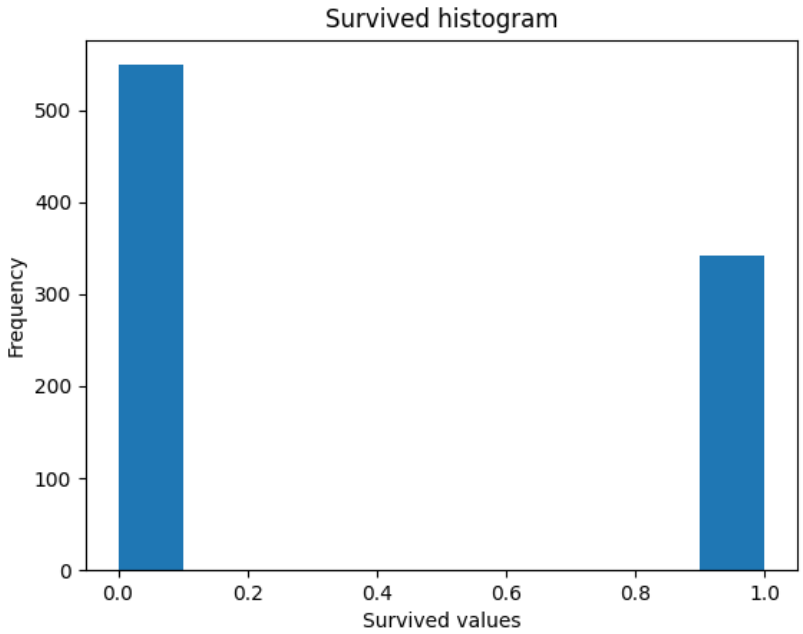

Figure 4: plot_Survived

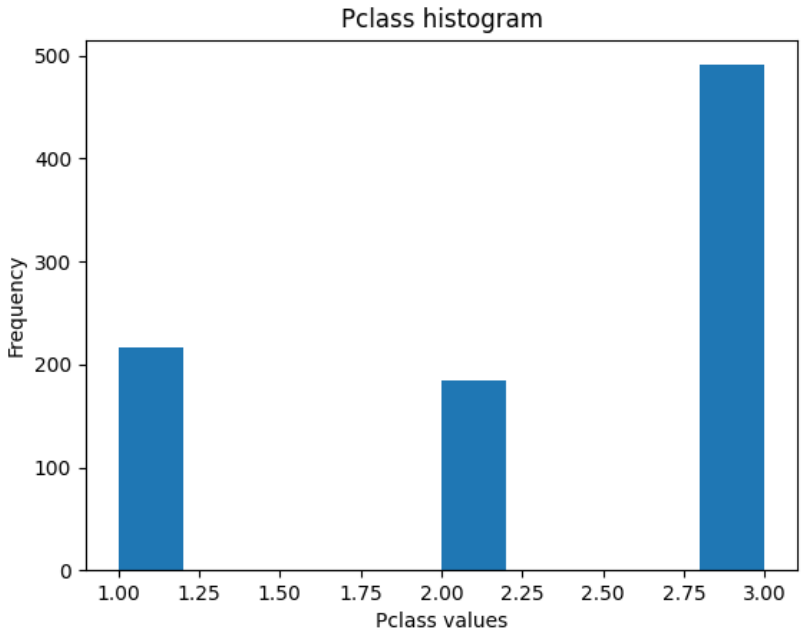

Figure 5: plot_Pclass

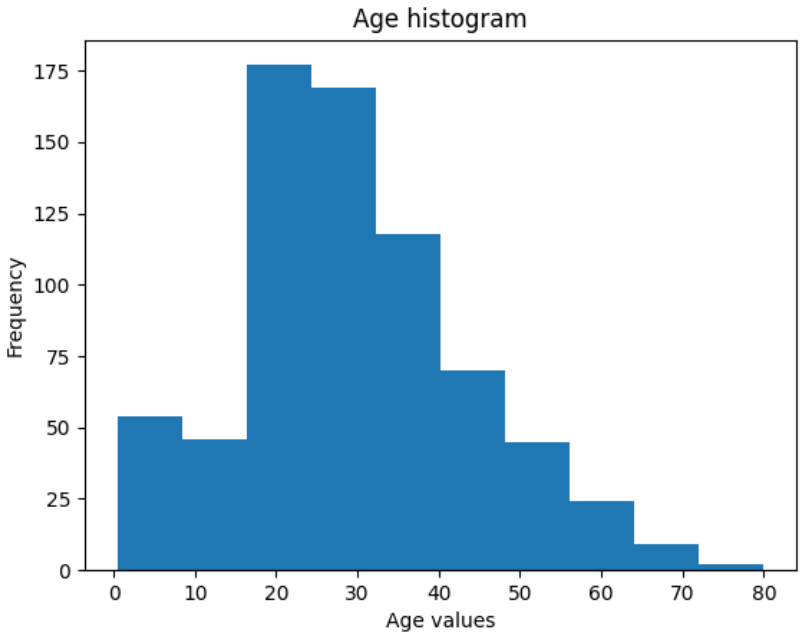

Figure 6: plot_Age

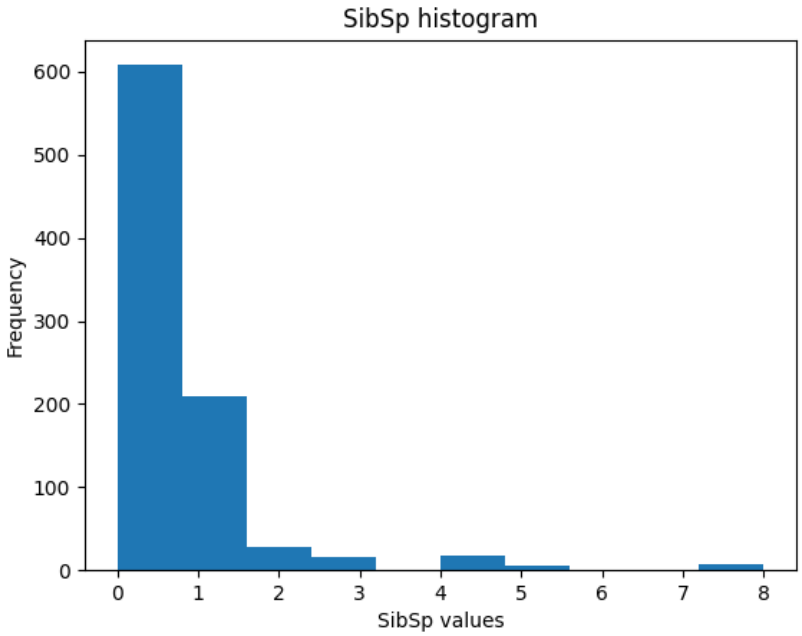

Figure 7: plot_SibSp

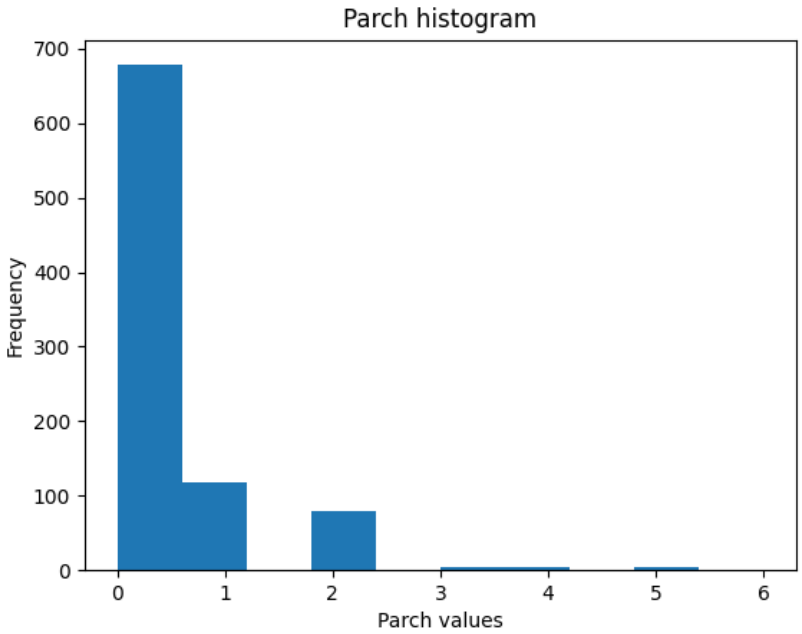

Figure 8: plot_Parch

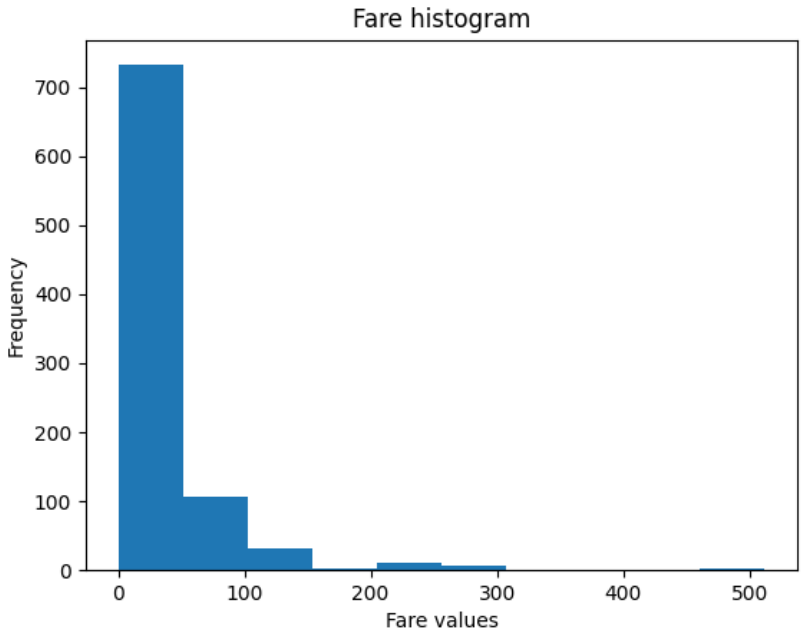

Figure 9: plot_Fare

Survived:

- Age: Numarul total este 52 Procentajul este 15.204678362573098%
- Cabin: Numarul total este 206 Procentajul este 60.23391812865497%
- Embarked: Numarul total este 2 Procentajul este 0.5847953216374269%

Not survived:

- Age: Numarul total este 125 Procentajul este 22.768670309653917%
- Cabin: Numarul total este 481 Procentajul este 87.61384335154827%

Cerinta 5

- Am adaugat o coloana in data frame, care sa ne spuna carui grup de varsta apartine un pasager, folosind functia apply. Folosind value_counts, am scos datele necesare pentru crearea unui pie-chart, care sa ilustreze distributia pasagerilor:

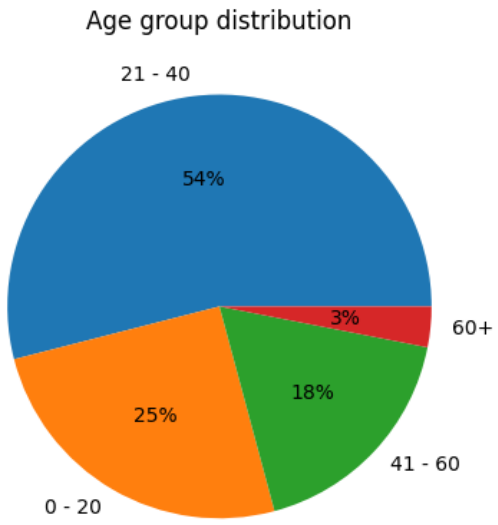

Figure 10: plot_AgeGroups

- De asemenea, am salvat noul data frame ca fisier csv.

Cerinta 6

- Folosind coloana creata la cerinta anterioara, am filtrat tabelul initial pentru a ramane doar cu barbatii, pe care i-am impartit dupa grupele de varsta si le-am calculat rata de supravietuire, pe care am pus-o in urmatorul grafic:

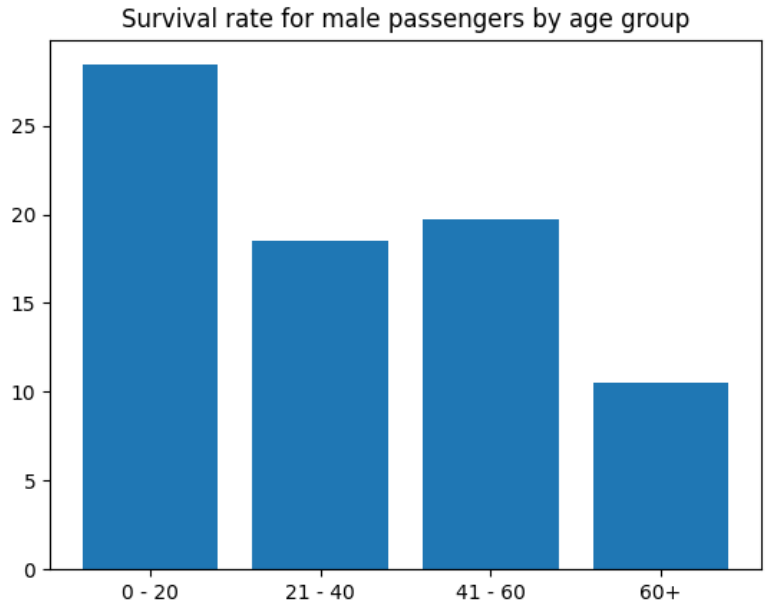

Figure 11: plot_Male_survival_rate

Cerinta 7

- Am filtrat din data frame-ul mare doar copiii (cu varsta sub 18 ani) si adultii (varsta peste 18 ani) si am calculat rata de supravietuire pentru ambele categorii:
- Am omis pasagerii a caror varsta este necunoscuta. De asemenea, am salvat cele doua data frame-uri create ca fisiere csv.

Cerinta 8

- Am impartit data frame-ul dupa cele doua clase date de coloana Survived. Folosind functia fillna(), am inlocuit valorile lipsa cu media celorlalte valori,

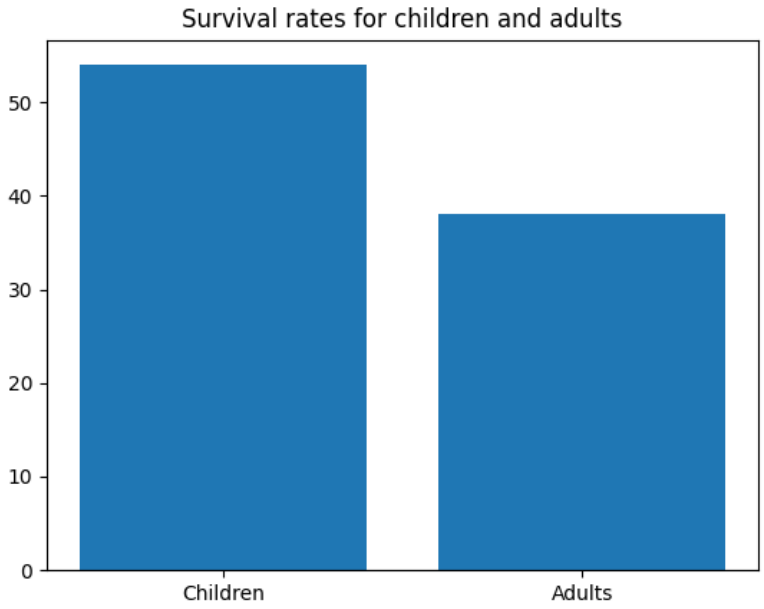

Figure 12: plot_Children&Adults_survival_rates

daca coloana continea date numerice, sau cu cea mai frecventa intrare, daca nu. Apoi, am inlocuit valorile calculate pe clase in data frame-ul original.

- Am salvat fisierul in format csv (filled.csv).

Cerinta 9

- Am identificat toate titlurile pe care le au pasagerii si le-am atribuit sexul corespunzator. Am reprezentat grafic frecventele titlurilor (vezi mai jos) si am calculat cate persoane au sexul diferit de cel presupus de titlu: o singura persoana (o persoana de sex feminin cu titlul “Dr”).

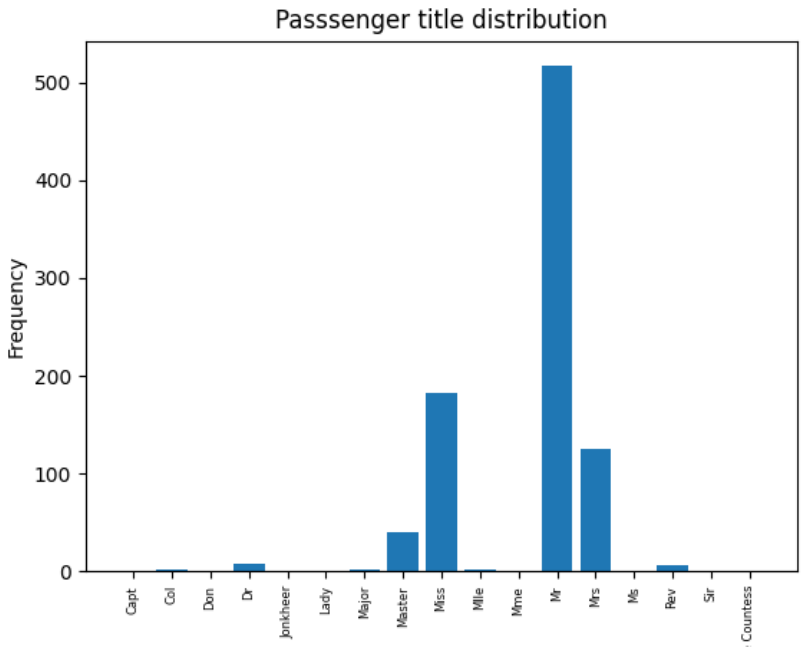

Figure 13: plot_Titles

- Am adaugat doua coloane noi la data frame: prima este titlul, iar a doua este sexul presupus de acesta. De aceea, am salvat noul data frame obtinut in format csv.

Cerinta 10

- Am adaugat o coloana noua data frame-ului, care sa spuna daca pasagerul a fost singur pe titanic (valorile de pe coloanele SibSp si Parch sunt 0) sau

nu. Am filtrat doar pasagerii singuri si am reprezentat intr-o histograma cati au supravietuit si cati nu:

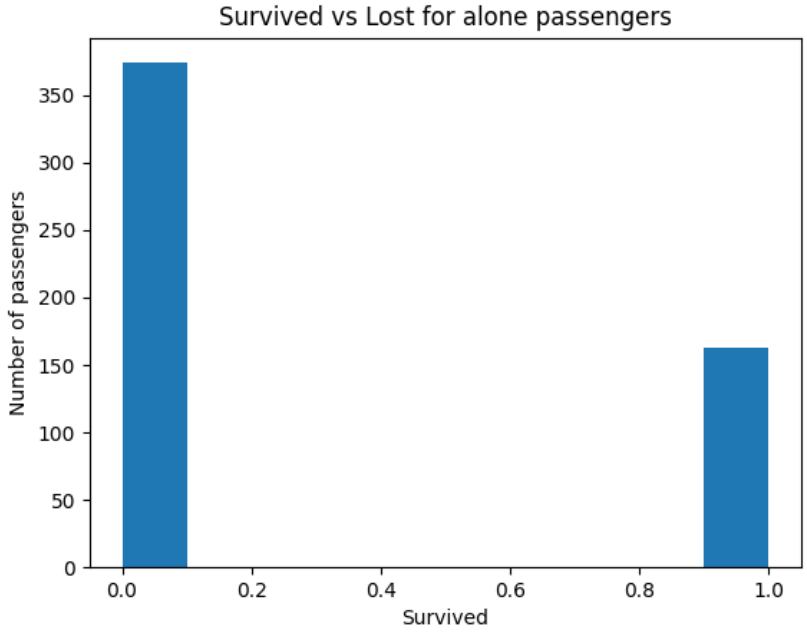

Figure 14: plot_Alone_survival_rate

- Considerand doar primii 100 pasageri, am reprezentat grafic relatia dintre clasa, pret si supravietuire:

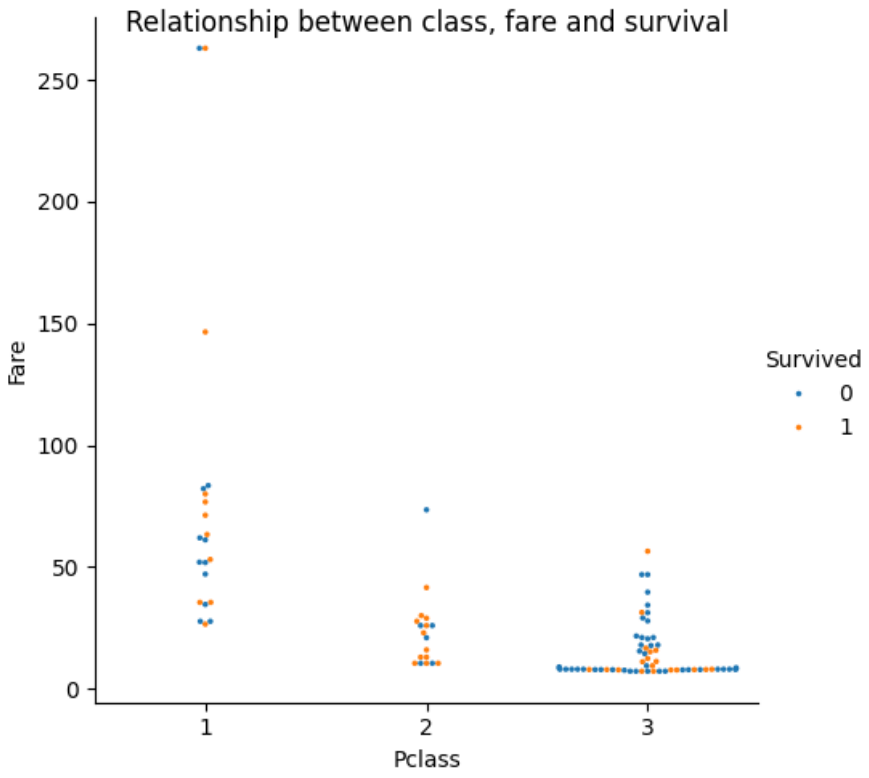

Figure 15: plot_Class-Fare-Survival_relationship