**МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ**

**Ордена Трудового Красного Знамени**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«Московский технический университет связи и информатики»**

Кафедра «Математическая Кибернетика и Информационные технологии»

Дисциплина «Введение в информационные технологии»

Лабораторная работа №3

«Анализ структуры и основных характеристик датасета пассажиров Титаника»

Выполнила:

Студентка группы БВТ2303

Морозова Ольга

**Цель работы:**

Изучить структуру и основные статистические показатели датасета пассажиров Титаника с использованием библиотеки matplotlib.

**Ход работы:**

**Задание 1.**

1. Определяем наличие пропущенных данных и заполняем их.

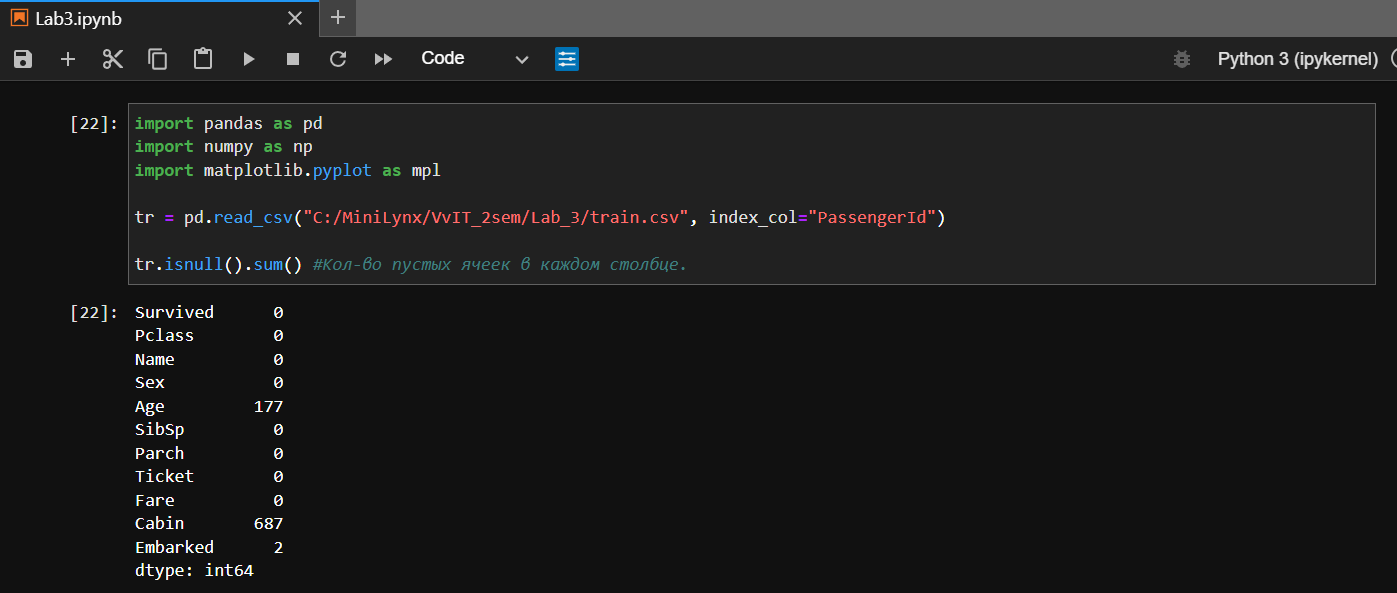
*import pandas as pd*

*import numpy as np*

*import matplotlib.pyplot as mpl*

*tr = pd.read\_csv("C:/MiniLynx/VvIT\_2sem/Lab\_3/train.csv", index\_col="PassengerId")*

*tr.isnull().sum() #Кол-во пустых ячеек в каждом столбце.*



*med = tr['Age'].median()*

*tr['Age'] = tr['Age'].fillna(value = med) #Заполним пустые ячейки возраста медианным значением.*

*emb = tr['Embarked'].value\_counts().idxmax()*

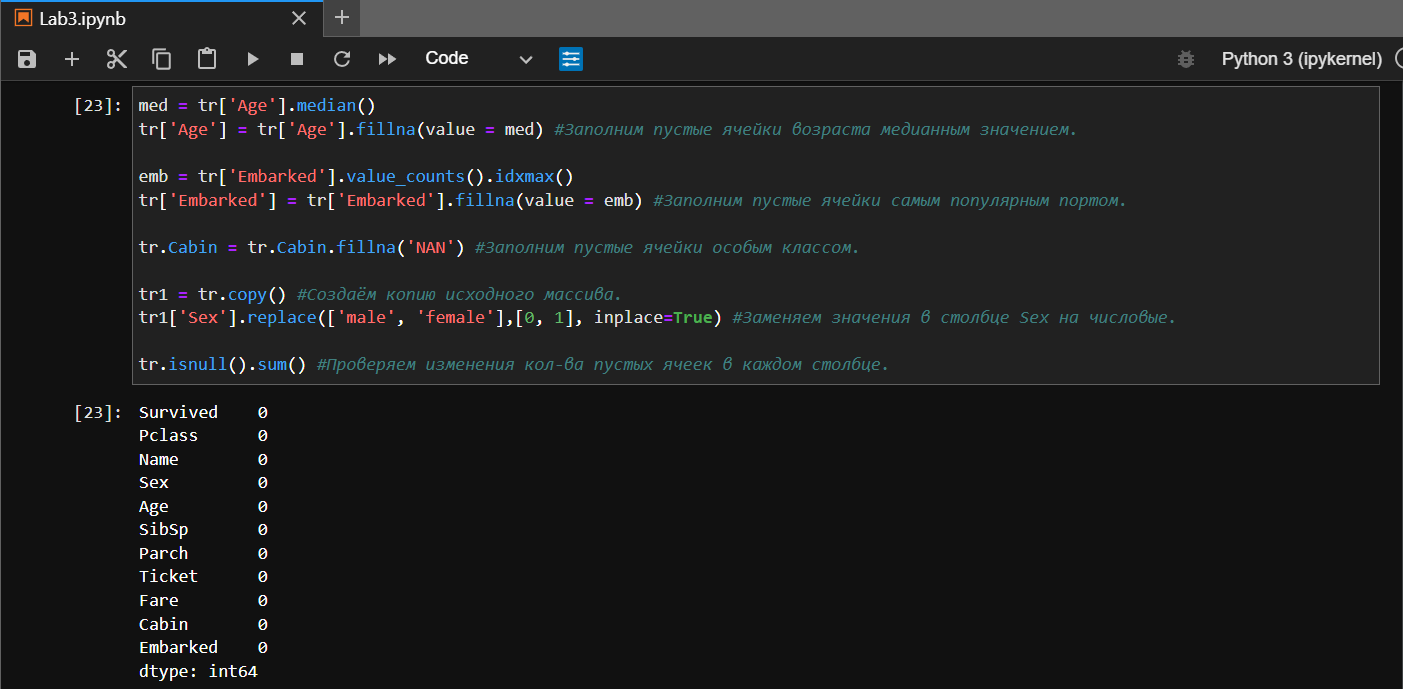
*tr['Embarked'] = tr['Embarked'].fillna(value = emb) #Заполним пустые ячейки самым популярным портом.*

*tr.Cabin = tr.Cabin.fillna('NAN') #Заполним пустые ячейки особым классом.*

*tr1 = tr.copy() #Создаём копию исходного массива.*

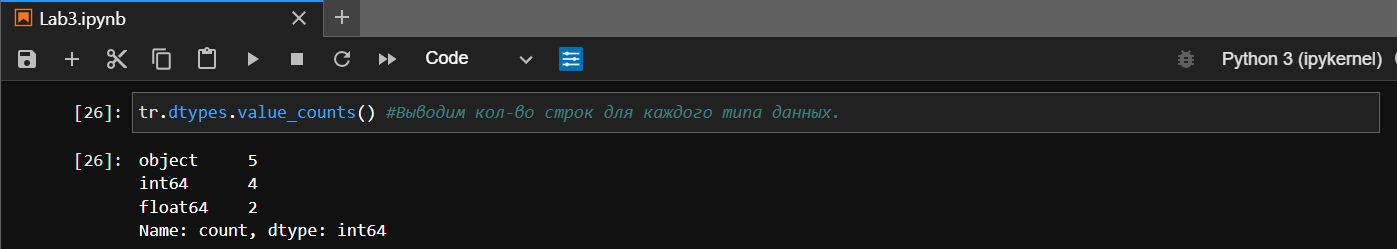
*tr1['Sex'].replace(['male', 'female'],[0, 1], inplace=True) #Заменяем значения в столбце Sex на числовые.*

*tr.isnull().sum() #Проверяем изменения кол-ва пустых ячеек в каждом столбце.*



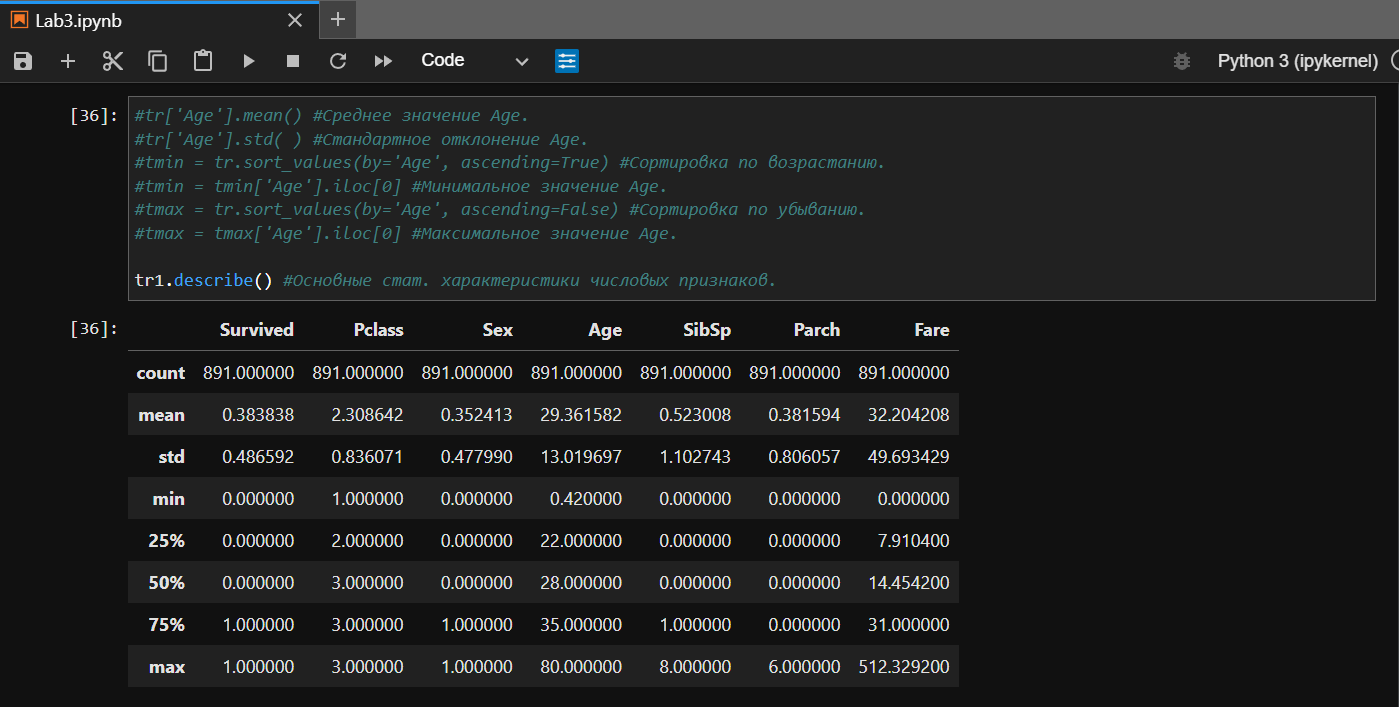
1. Определяем количество признаков и их типы.

*tr.dtypes.value\_counts() #Выводим кол-во строк для каждого типа данных.*



1. Выводим основные статистические характеристики (среднее значение, стандартное отклонение, минимальное и максимальное значения) для числовых признаков.

*tr1.describe() #Основные стат. характеристики числовых признаков.*



Альтернативные способы получить эти данные:

*tr['Age'].mean() #Среднее значение Age.*

*tr['Age'].std( ) #Стандартное отклонение Age.*

*tmin = tr.sort\_values(by='Age', ascending=True) #Сортировка по возрастанию.*

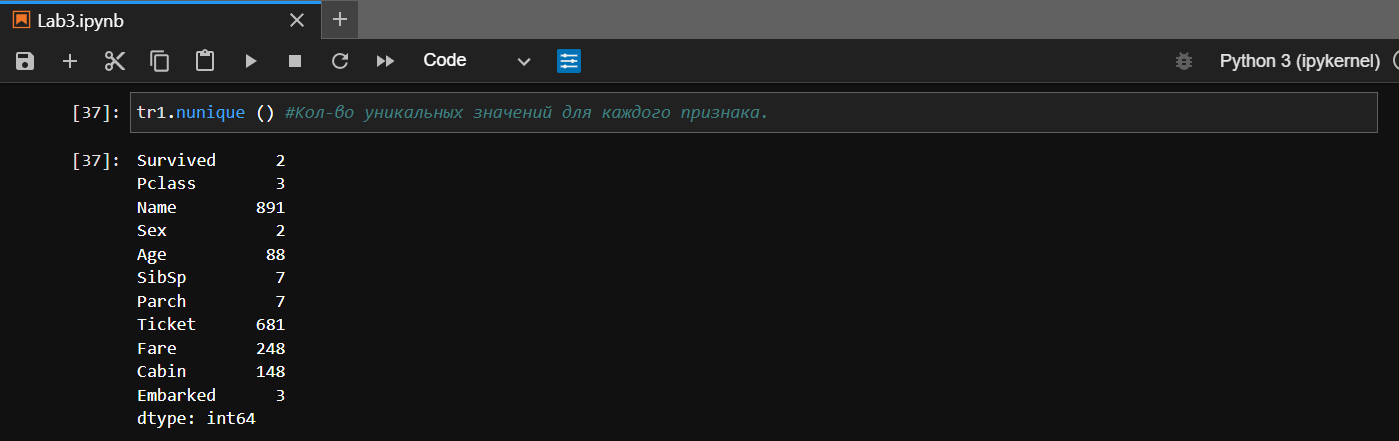
*tmin = tmin['Age'].iloc[0] #Минимальное значение Age.*

*tmax = tr.sort\_values(by='Age', ascending=False) #Сортировка по убыванию.*

*tmax = tmax['Age'].iloc[0] #Максимальное значение Age.*

1. Выводим уникальные значения для категориальных признаков в датасете.

*tr1.nunique () #Кол-во уникальных значений для каждого признака.*

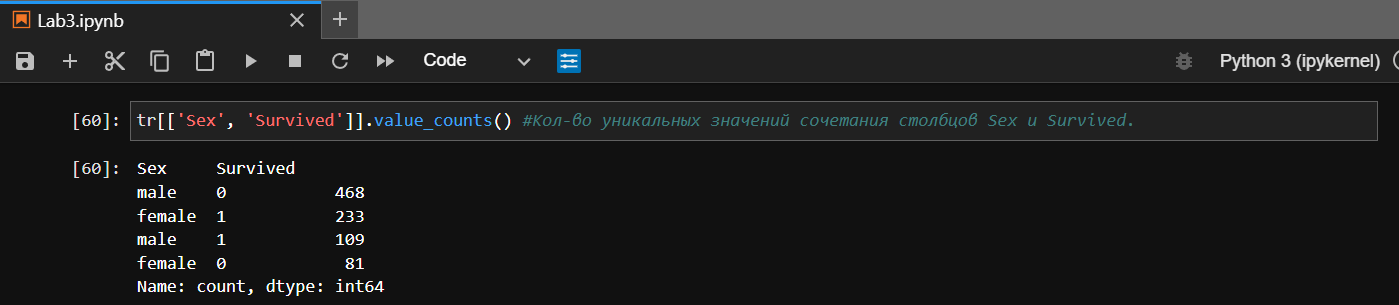


**Задание 2.**

Построение графика "Распределение выживших по полу" с использованием библиотеки matplotlib.

1. Столбчатая:

*tr[['Sex', 'Survived']].value\_counts() #Кол-во уникальных значений сочетания столбцов Sex и Survived.*



*sex = np.array(tr['Sex'].value\_counts()) #Делаем массив NumPy из кол-ва мужчин и женщин.*

*s = ('male', 'female')*

*m0, f1, m1, f0 = tr[['Sex', 'Survived']].value\_counts() #Записываем полученные ранее значения в 4 переменные.*

*#Формируем из них массив:*

*surv = {*

*'Погибло': (m0, f0),*

*'Выжило': (m1, f1),*

*}*

*#Задаём необходимые переменные для параметров графика:*

*x = np.arange(len(sex))*

*multiplier = 0*

*fig, ax = mpl.subplots(layout='constrained')*

*#Создаём несколько графиков, накладывающихся один на другой:*

*for atr, meas in surv.items():*

*offset = width \* multiplier*

*rects = ax.bar(x + offset, meas, width = 0.25, label=atr)*

*ax.bar\_label(rects, padding=3)*

*multiplier += 1*

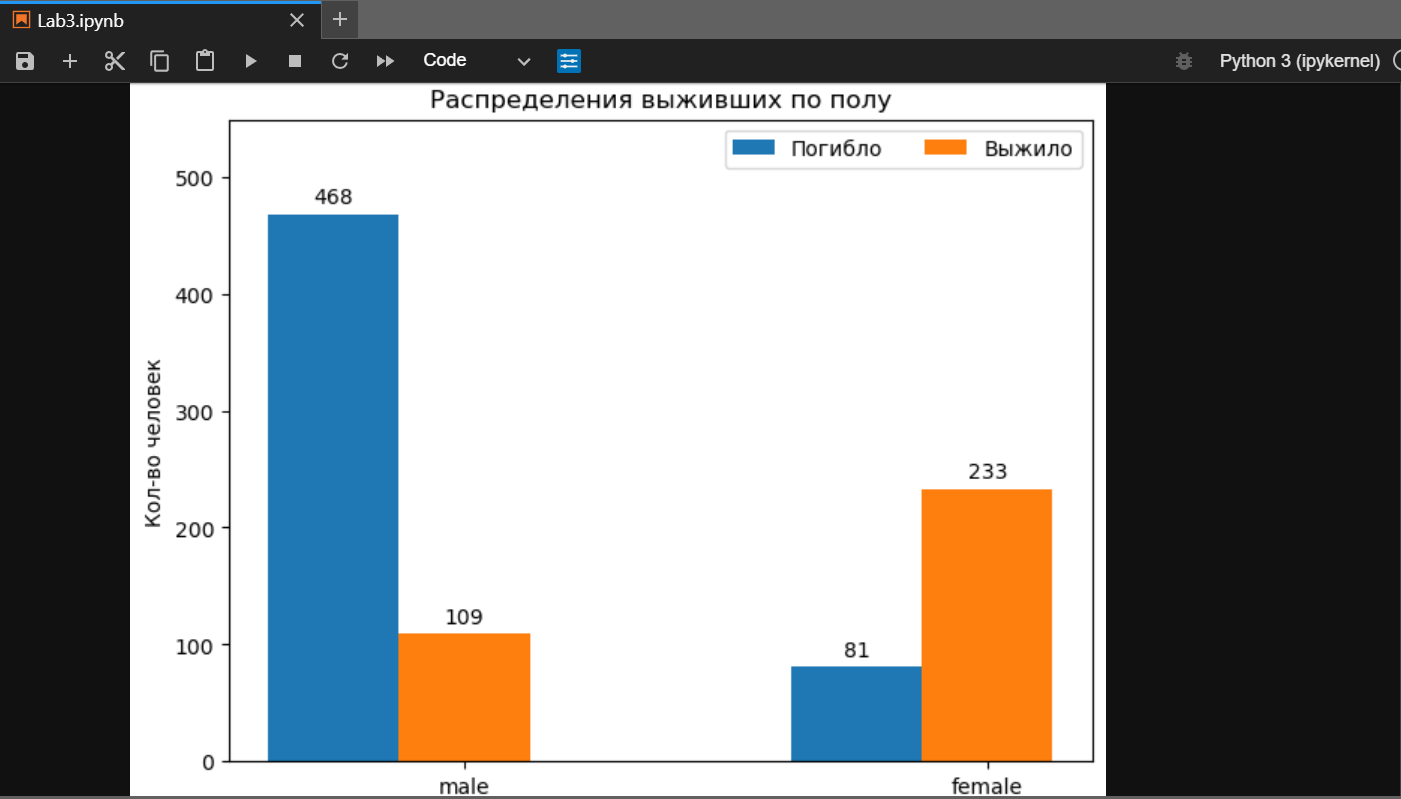
*ax.set\_ylabel('Кол-во человек')*

*ax.set\_title('Распределения выживших по полу')*

*ax.set\_xticks(x + width, s)*

*ax.legend(loc='upper right', ncols=2)*

*ax.set\_ylim(0, m0 + f0)*



1. Круговая:

*#Мужчины.*

*mpl.figure(0)*

*m = (tr['Sex'] == 'male').sum()*

*m1 = len(tr.loc[((tr['Sex'] == 'male') & (tr['Survived'] == 1))])*

*m0 = m - m1*

*mpl.pie([m0, m1], labels=["Погибло", 'Выжило'], autopct='%1.0f%%', colors=('silver', 'cornflowerblue'))*

*mpl.title('male')*

*#Женщины.*

*mpl.figure(1)*

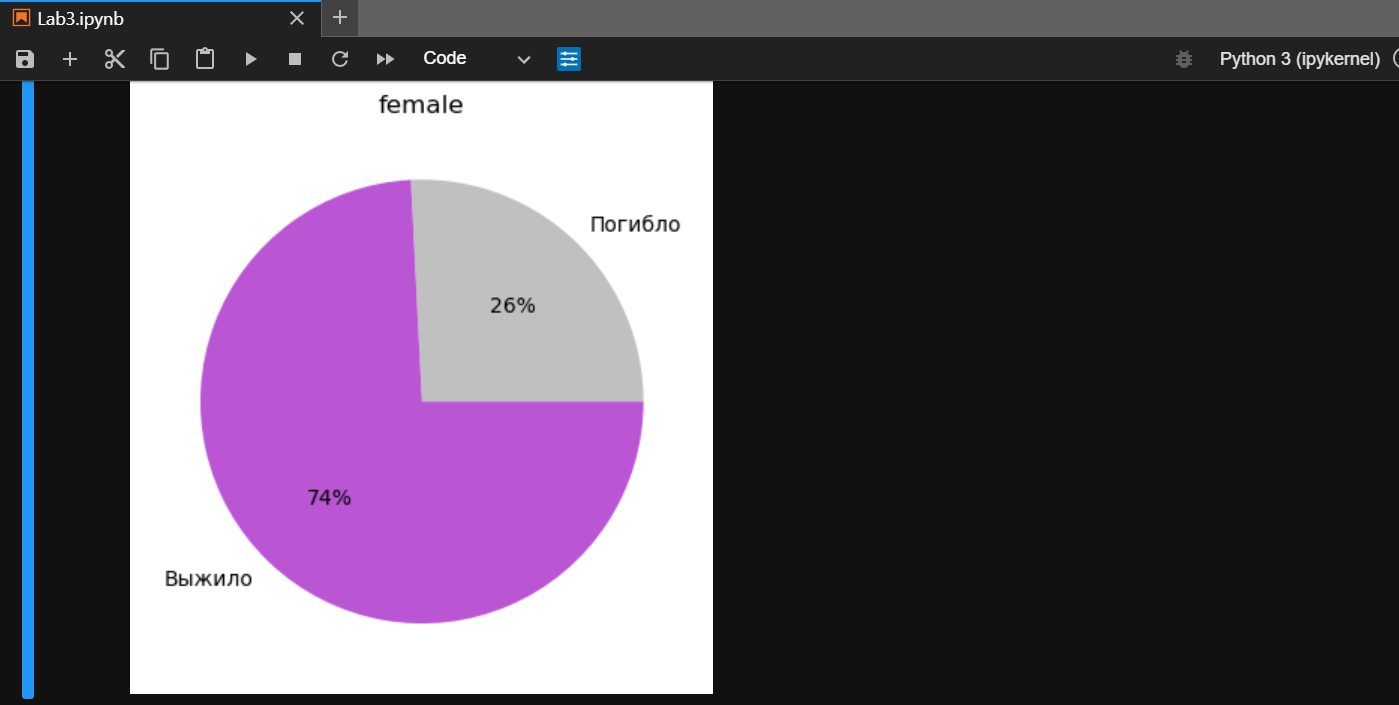
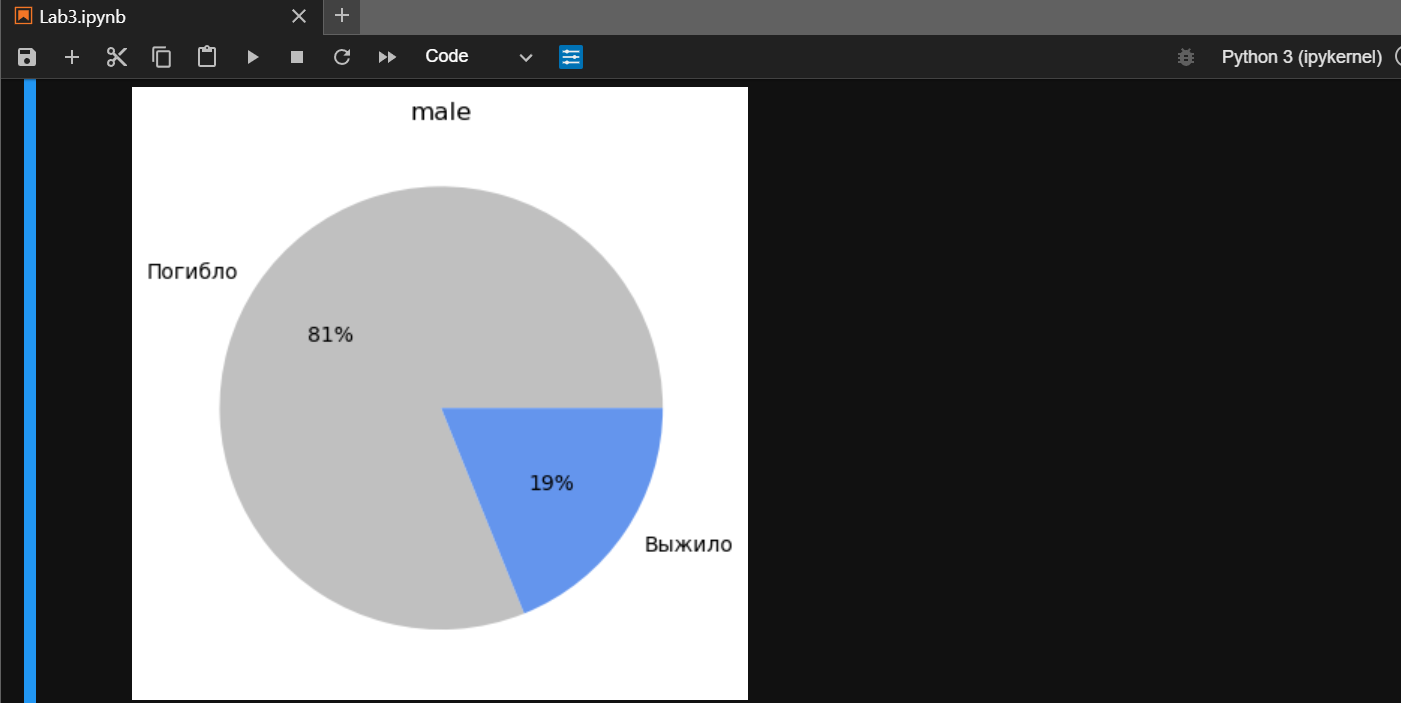
*f = (tr['Sex'] == 'female').sum()*

*f1 = len(tr.loc[((tr['Sex'] == 'female') & (tr['Survived'] == 1))])*

*f0 = f - f1*

*mpl.pie([f0, f1], labels=["Погибло", 'Выжило'], autopct='%1.0f%%', colors=('silver', 'mediumorchid'))*

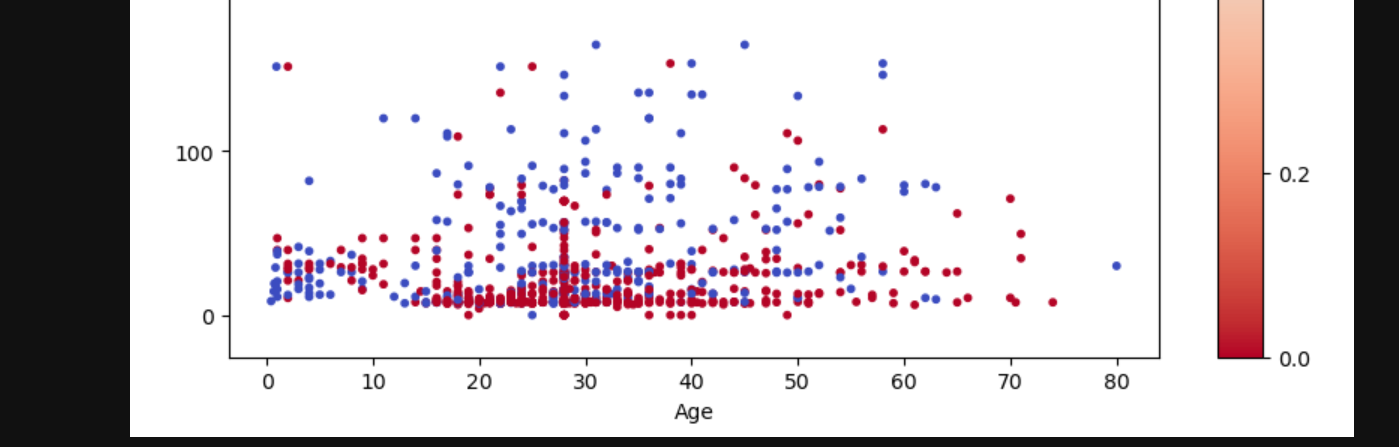
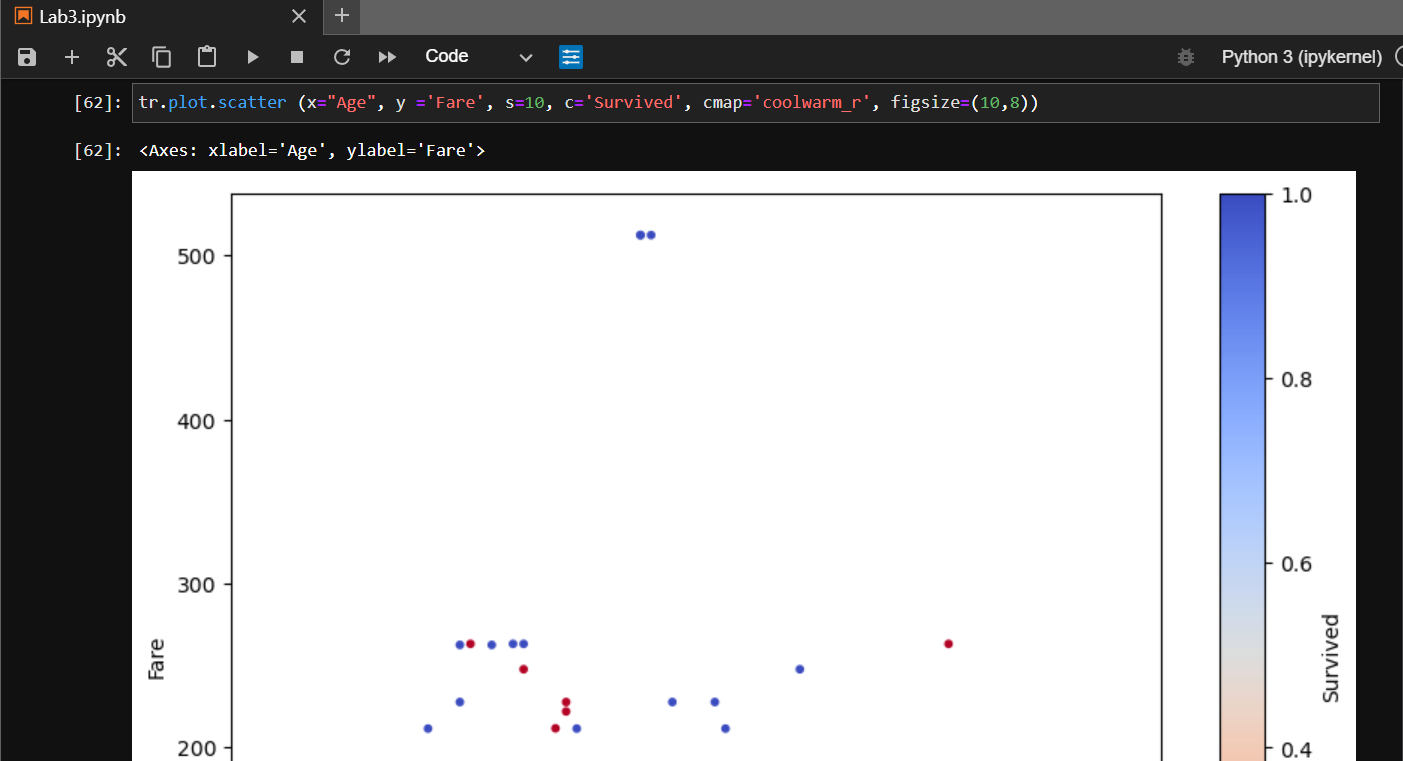
*mpl.title('female')*



**Задание 3.**

Создать scatter plot для анализа взаимосвязи между возрастом и ценой билета.

*tr.plot.scatter (x="Age", y ='Fare', s=10, c='Survived', cmap='coolwarm\_r', figsize=(10,8))*



**Вывод:**

Мы изучили структуру и основные статистические показатели датасета пассажиров Титаника, научились использовать библиотеку matplotlib для вывода этих данных в наглядном виде (диаграммы).