Зозуля О.А. ИУ5-25М РК1

# **Вариант 6**

***Задача №6.***

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения средним значением.

Датасет: Бар

import matplotlib  
from mpl\_toolkits import mplot3d  
import pandas as pd  
from matplotlib import pyplot as plt  
import numpy as np  
import seaborn as sns  
import os

#Загрузка данных из файла

path=os.environ["userprofile"]+"\\"+".atom"+"\\"+"Transactions.csv"  
print(path)

data = pd.read\_csv(path)  
print(data)

#Замена пропущенных значений средним значением

mean\_value = data['Percentage\_markup'].mean()  
data['Percentage\_markup'].fillna(mean\_value, inplace=True)

#Построение парных диаграмм

sns.pairplot(data, vars=['Percentage\_markup', 'Profit'], diag\_kind='kde')

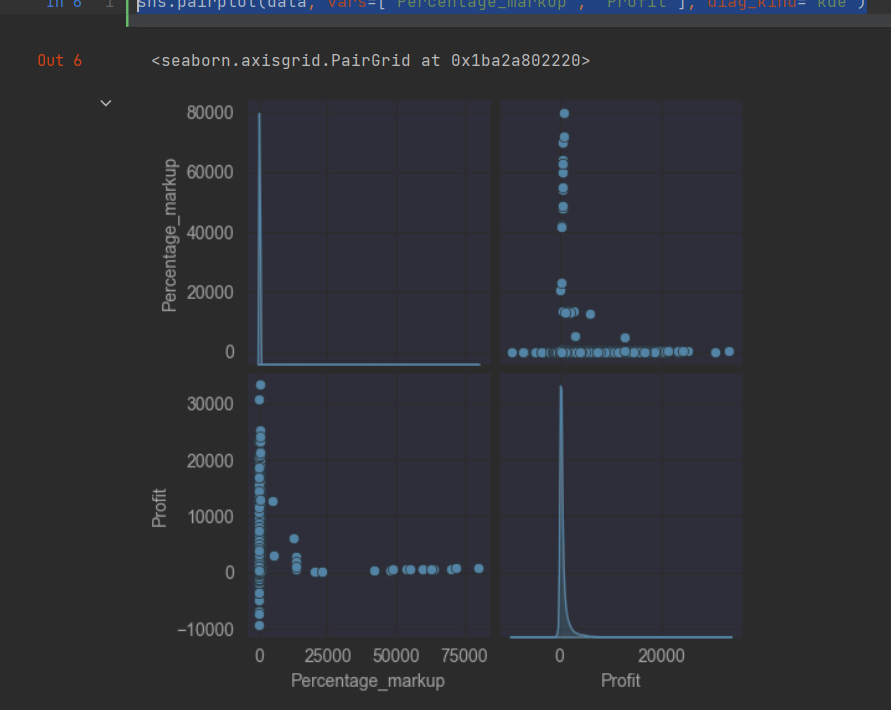


Рисунок 1. Парная диаграмма (Процентная отметка и Прибыль)

#Сохранение изменённых данных в файл

data.to\_csv('file\_with\_filled\_values.csv', index=False)

**Вывод:**

В этом примере мы загружаем данные из файла ‘Transactions.csv’, заменяем пропущенные значения в столбце 'Percentage\_markup' средним значением, строим парные диаграммы для столбцов 'Percentage\_markup' и ‘Profit’ с помощью метода pairplot из библиотеки seaborn и сохраняем измененные данные в файл 'file\_with\_filled\_values.csv'.

Я использовала параметр diag\_kind='kde' для построения графиков плотности распределения на диагонали парных диаграмм.

***Задача №26.***

Для набора данных для одного (произвольного) числового признака проведите обнаружение и замену (найденными верхними и нижними границами) выбросов на основе правила трех сигм.

Датасет: Об убийствах

Вот пример кода на языке Python, который демонстрирует, как это можно сделать:

import matplotlib  
from mpl\_toolkits import mplot3d  
import pandas as pd  
from matplotlib import pyplot as plt  
import numpy as np  
import seaborn as sns  
import os

#Загрузка данных из файла

path=os.environ["userprofile"]+"\\"+".atom"+"\\"+"database.csv"  
print(path)

data = pd.read\_csv(path)  
print(data)

#Вычисление среднего и стандартного отклонения

mean\_value = data['Year']. mean ()  
std\_value = data['Year']. std()

#Вычисление верхней и нижней границы

lower\_bound = mean\_value - 3 \* std\_value  
upper\_bound = mean\_value + 3 \* std\_value

#Замена выбросов на границы

data.loc[data['Year'] < lower\_bound, 'Year'] = lower\_bound  
data.loc[data['Year'] > upper\_bound, 'Year'] = upper\_bound

#Построение парных диаграмм

sns.pairplot(data, vars=['Year', 'Incident'][:500], diag\_kind='kde')

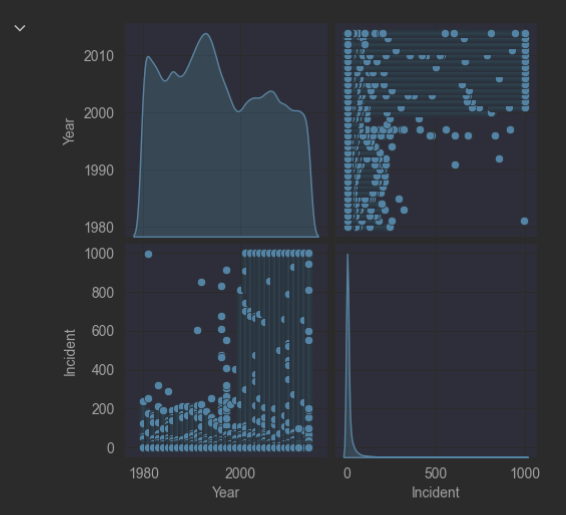


Рисунок 2. Парная диаграмма (Год и количество инцидентов)

**Вывод:**

В этом примере мы загружаем данные из файла ‘database.csv’, вычисляем среднее значение и стандартное отклонение для столбца ‘Year’, вычисляем верхнюю и нижнюю границы на основе правила трех сигм, заменяем выбросы на границы, строим парные диаграммы для столбцов ‘Year’ и ‘Incident’ с помощью метода pairplot из библиотеки seaborn и сохраняем измененные данные в файл 'file\_with\_replaced\_outliers.csv'.

Я использовала параметр diag\_kind='kde' для построения графиков плотности распределения на диагонали парных диаграмм.