МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

<<ЛЬВІВСЬКА ПОЛІТЕХНІКА>>

Інститут ІКНІ

Кафедра систем штучного інтелекту



ЗВІТ

Лабораторна робота **№4**

З курсу “ Інтелектуальний аналіз даних”

Виконав:

**Гавриляк Тарас**

гр. **КН-310**

Прийняв(ла):

Якимишин Х.М.

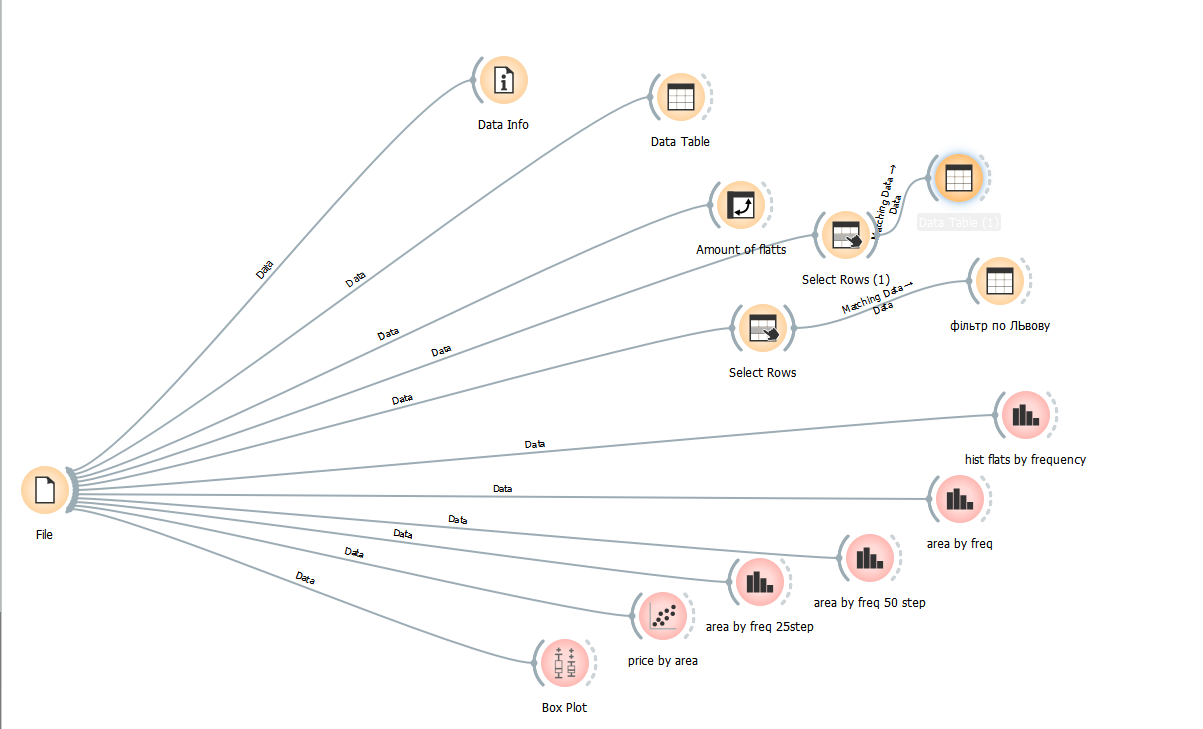
Львів – 2020

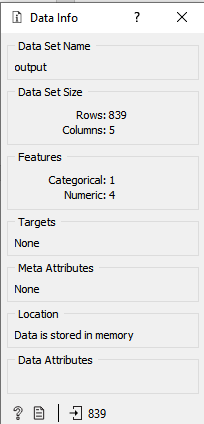
**Лабораторна №4**

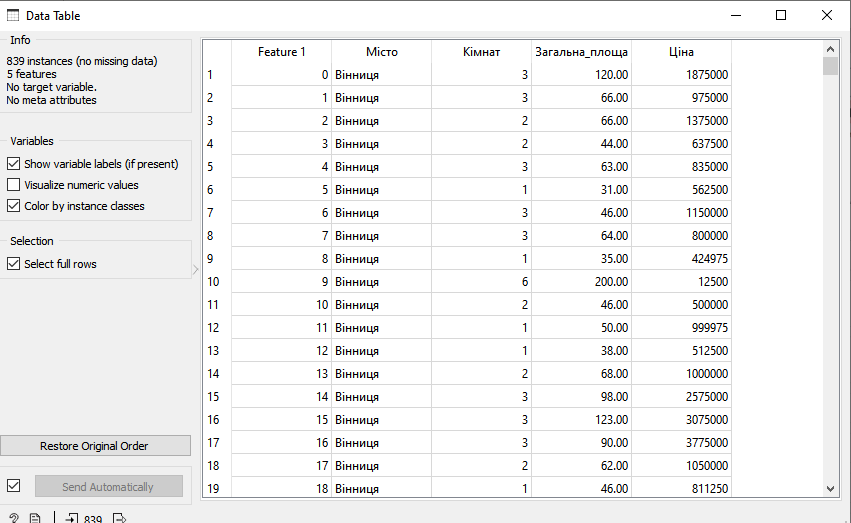
**Візуальний аналіз даних методами Orange**

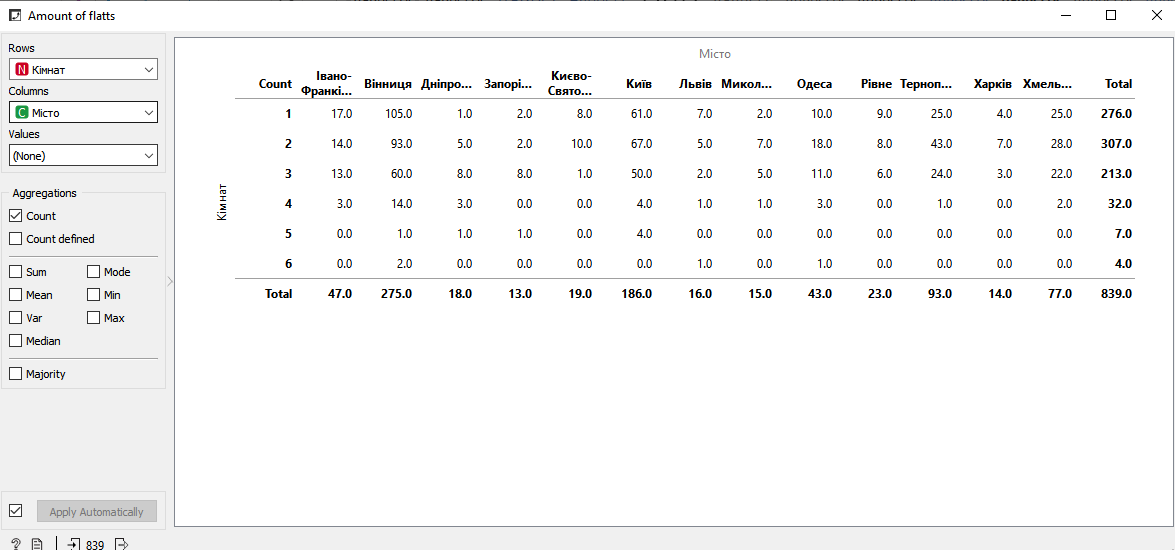
1. Використовуючи Orange повторити розрахунки лабораторних робіт 1 та 2 з предмету візуалізація даних.
2. Оформити звіт з таблицею переваг та недоліків R/Pyhton vs Orange

**Завдання до лабораторної №1 з візуалізації.**

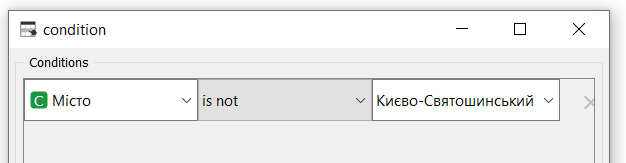
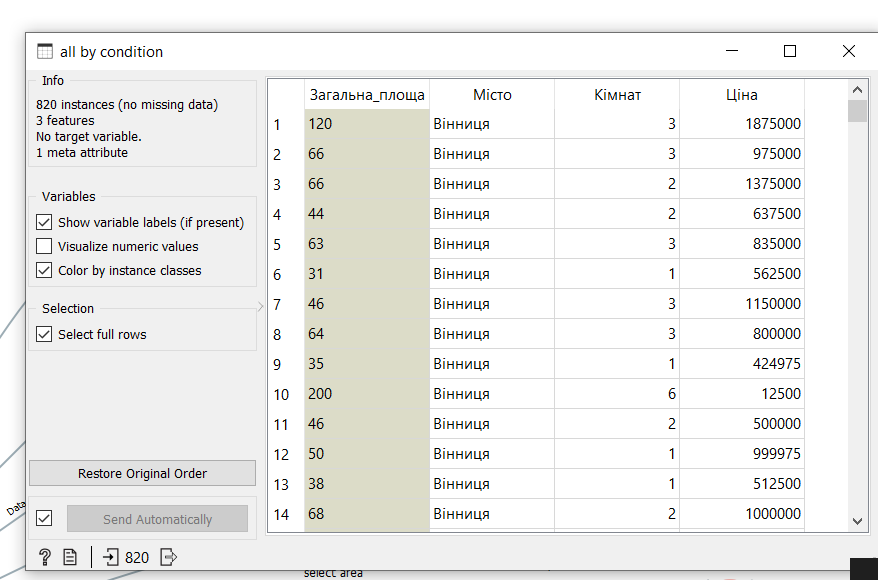
****

****

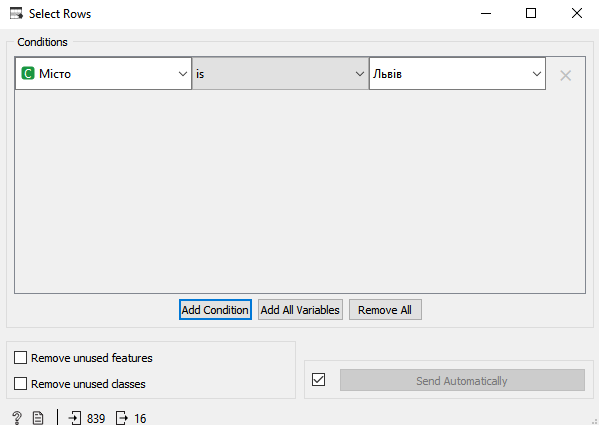
****

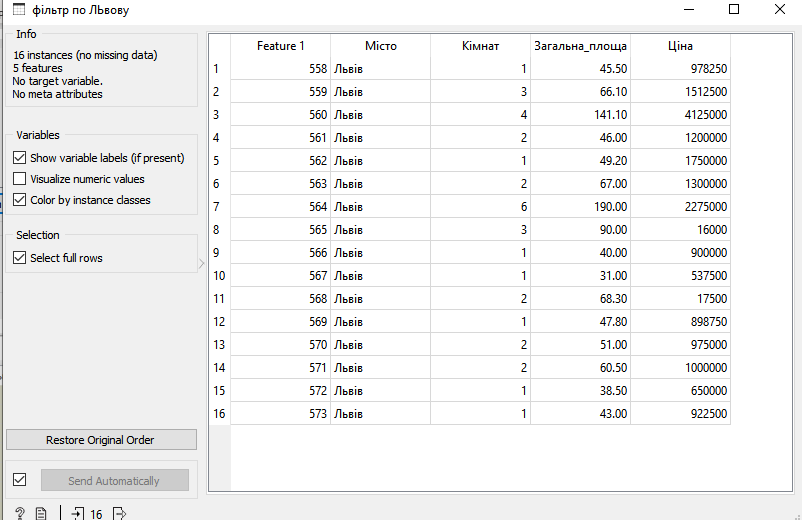
****

Всі міста окрім Києво-Святошинського району

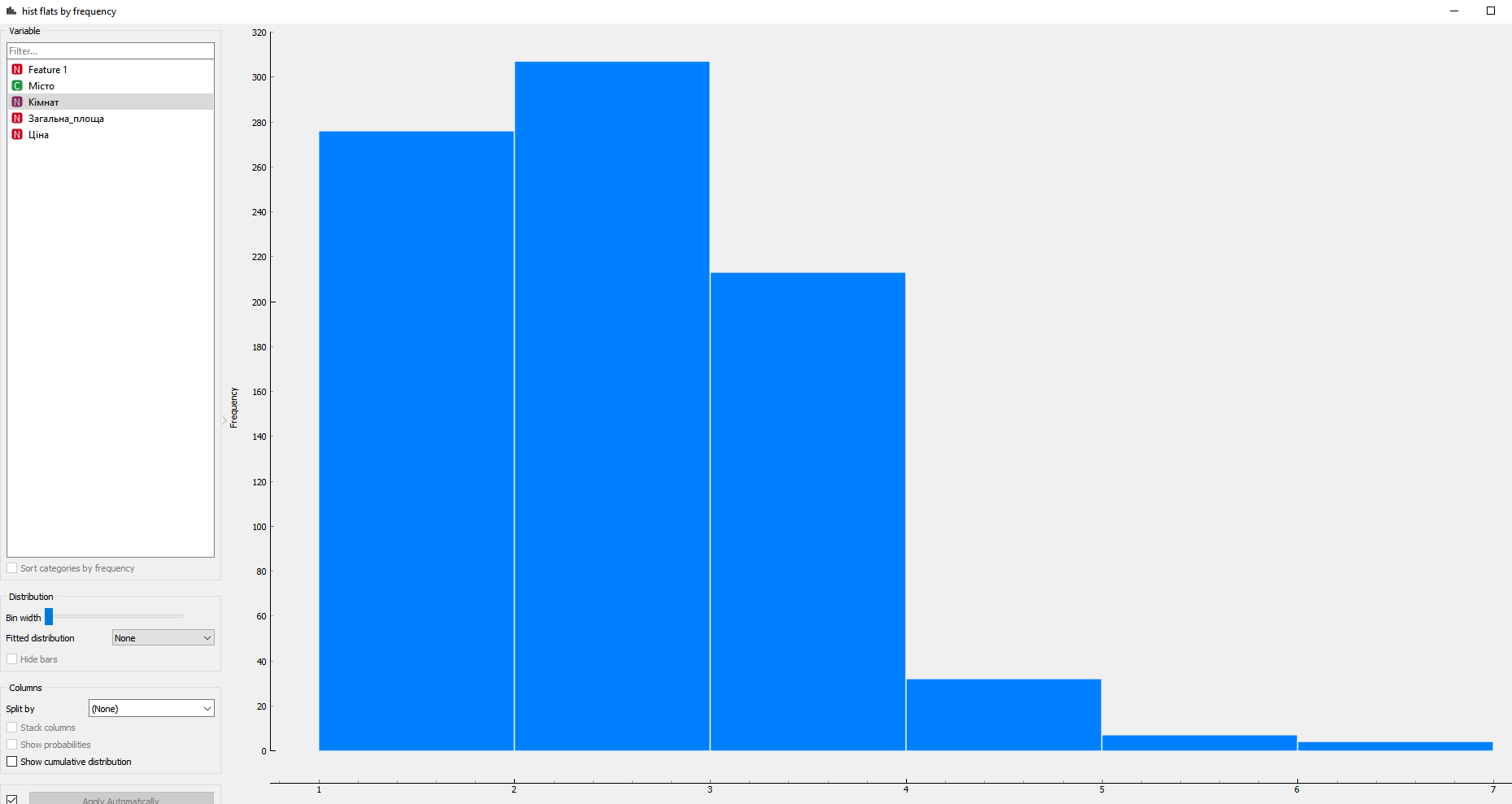
**** ****

Вибірка по конкретному місту

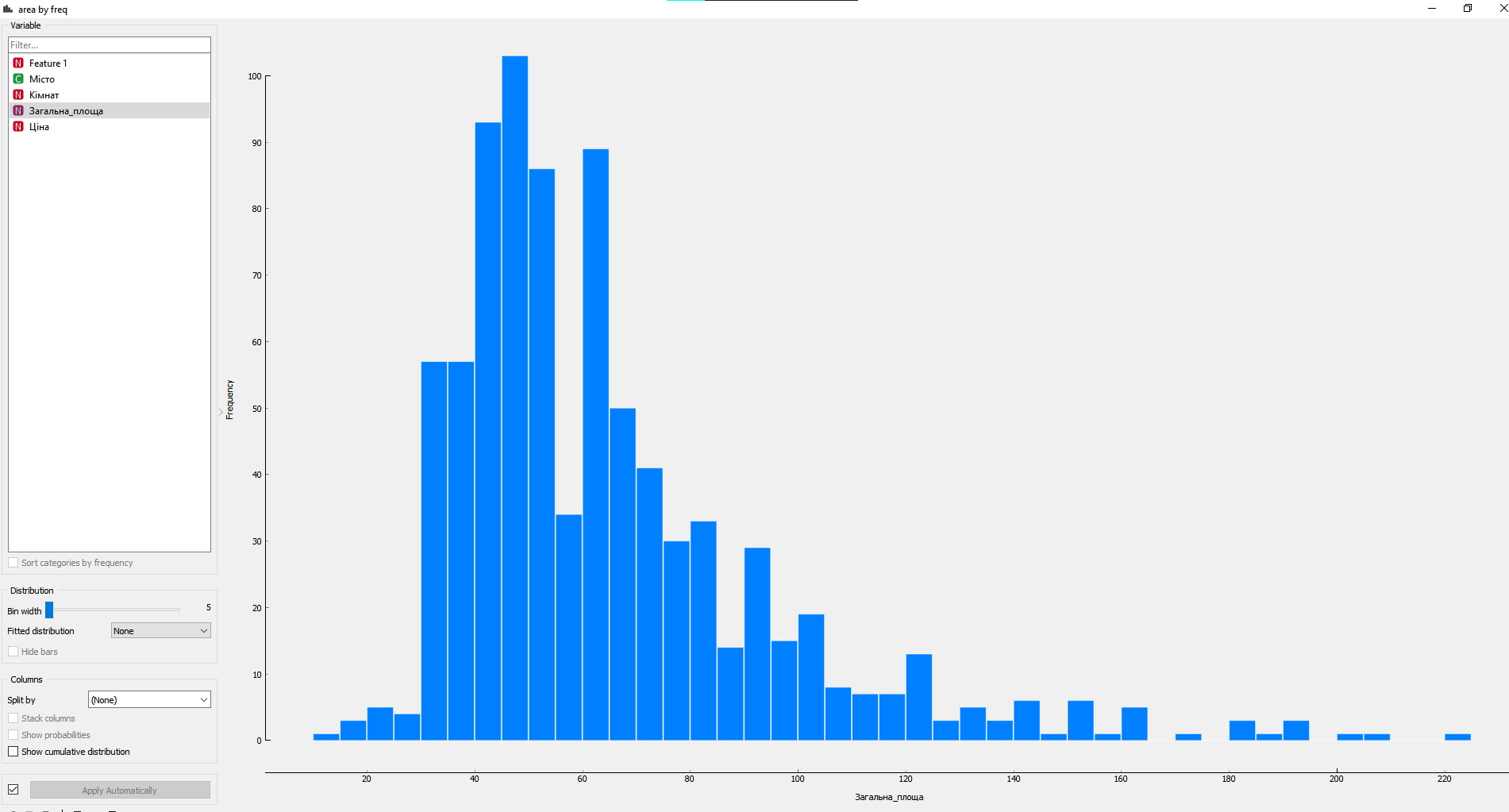
****

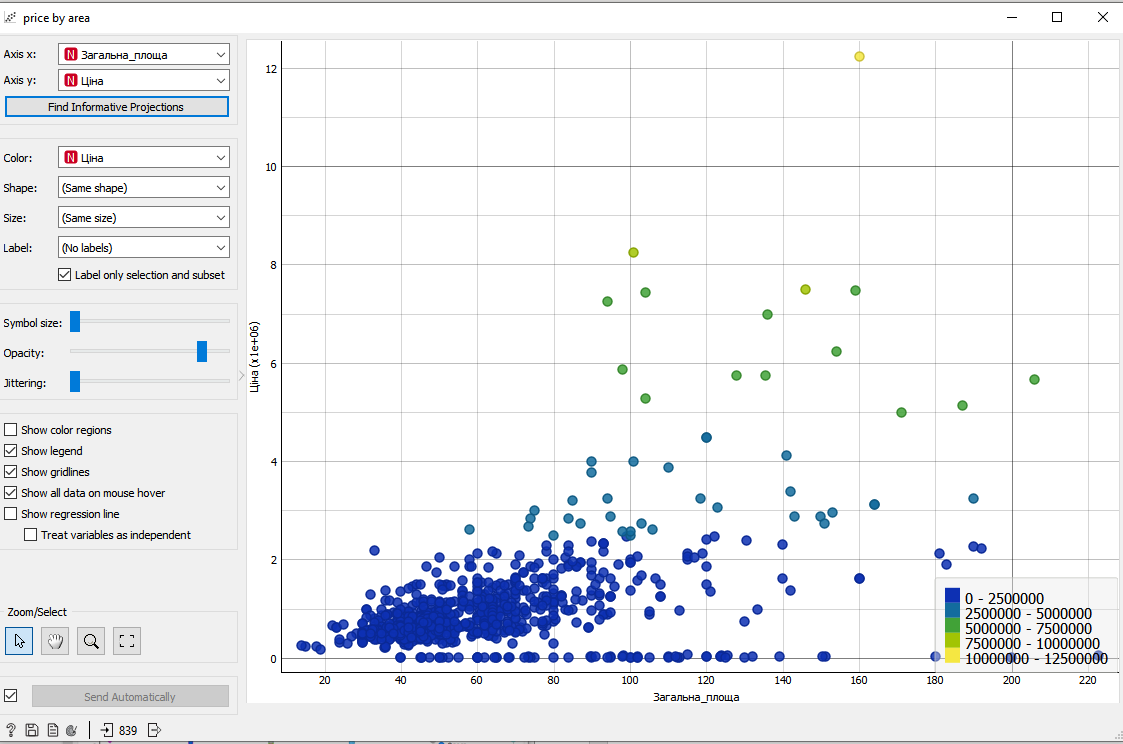
****

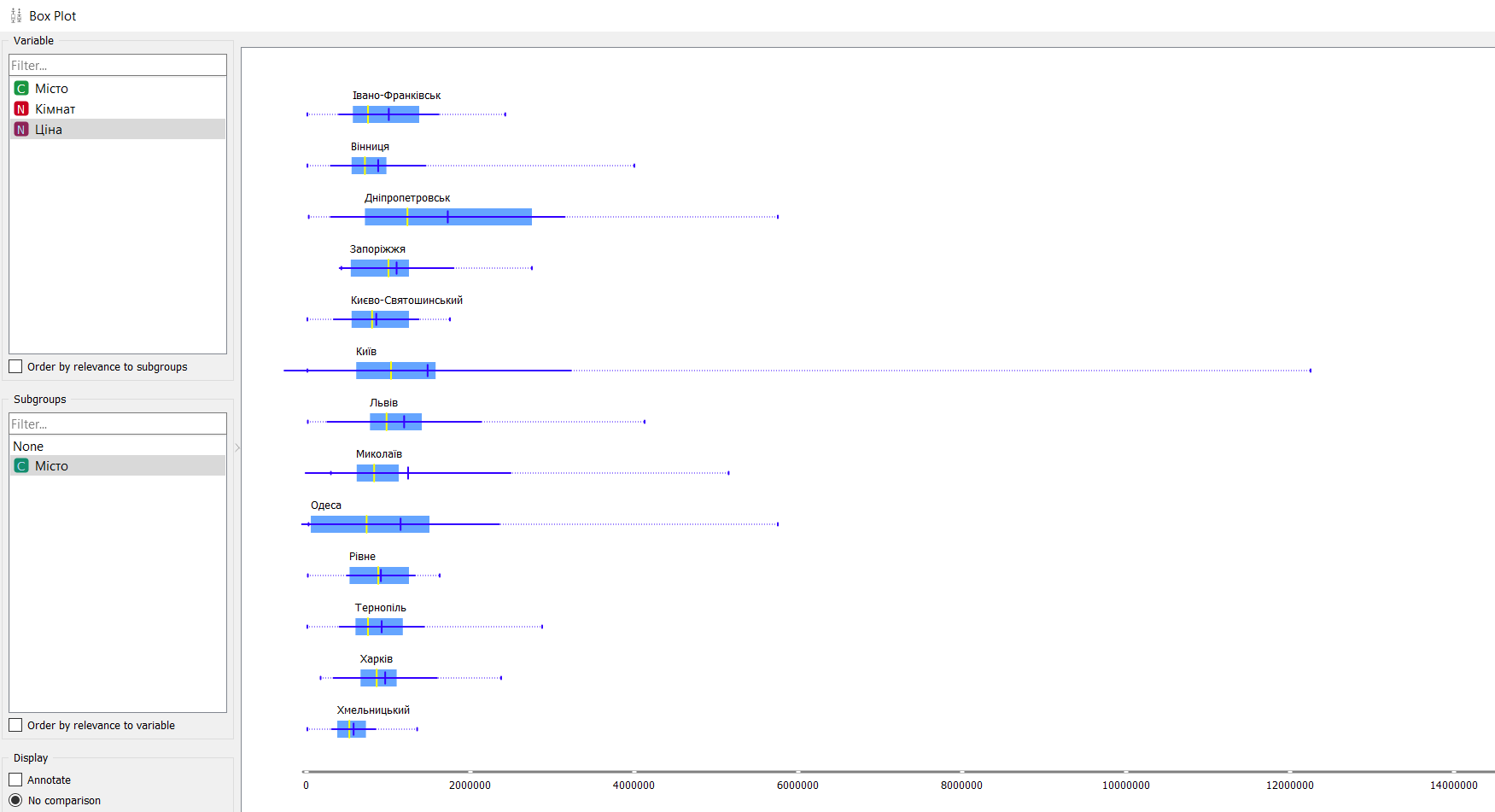
Гістограма (кількість оголошень по кількості кімнат)



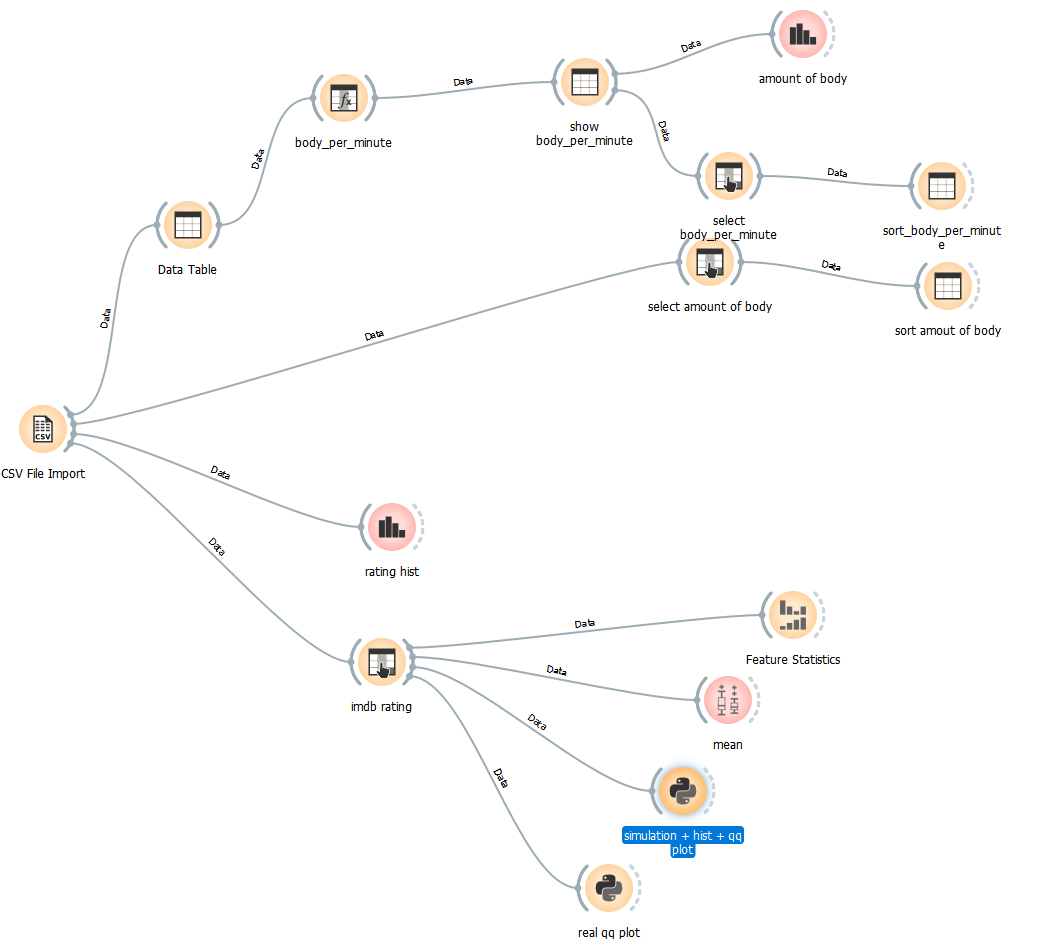
Розподіл оголошень по цінах



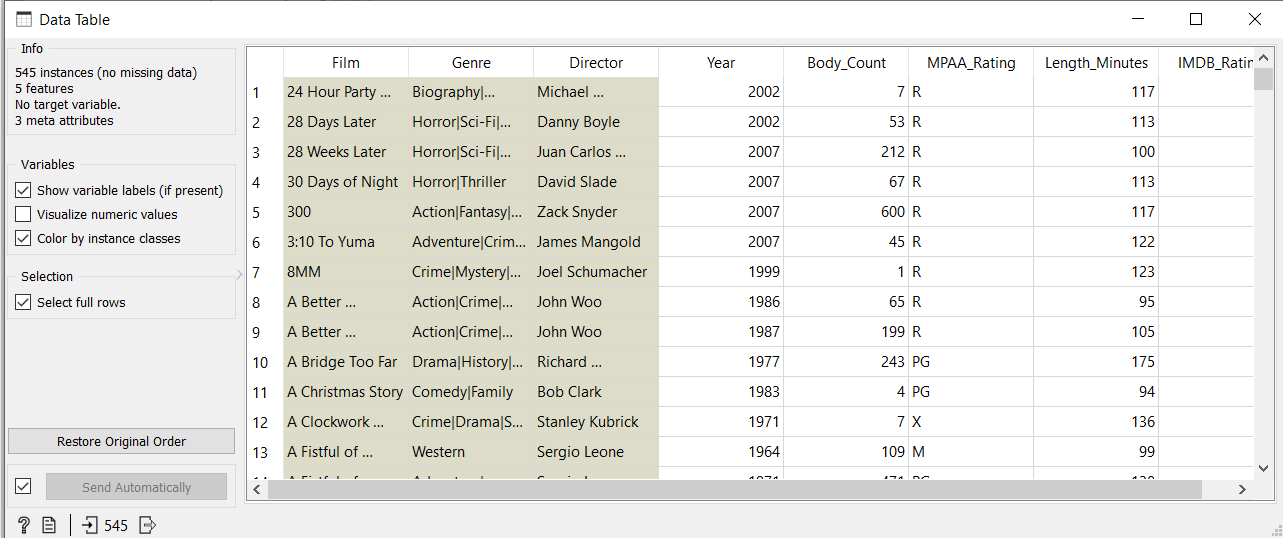




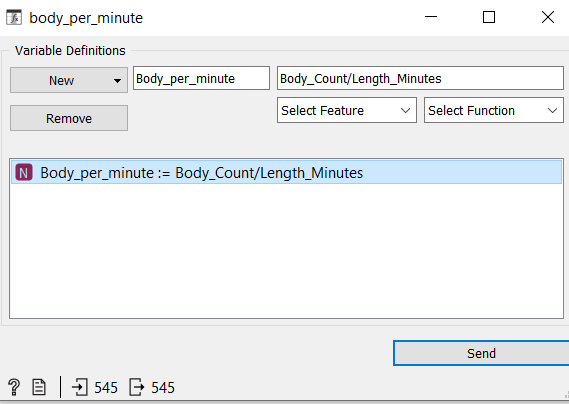
**Завдання до лабораторної №2 з візуалізації.**

****

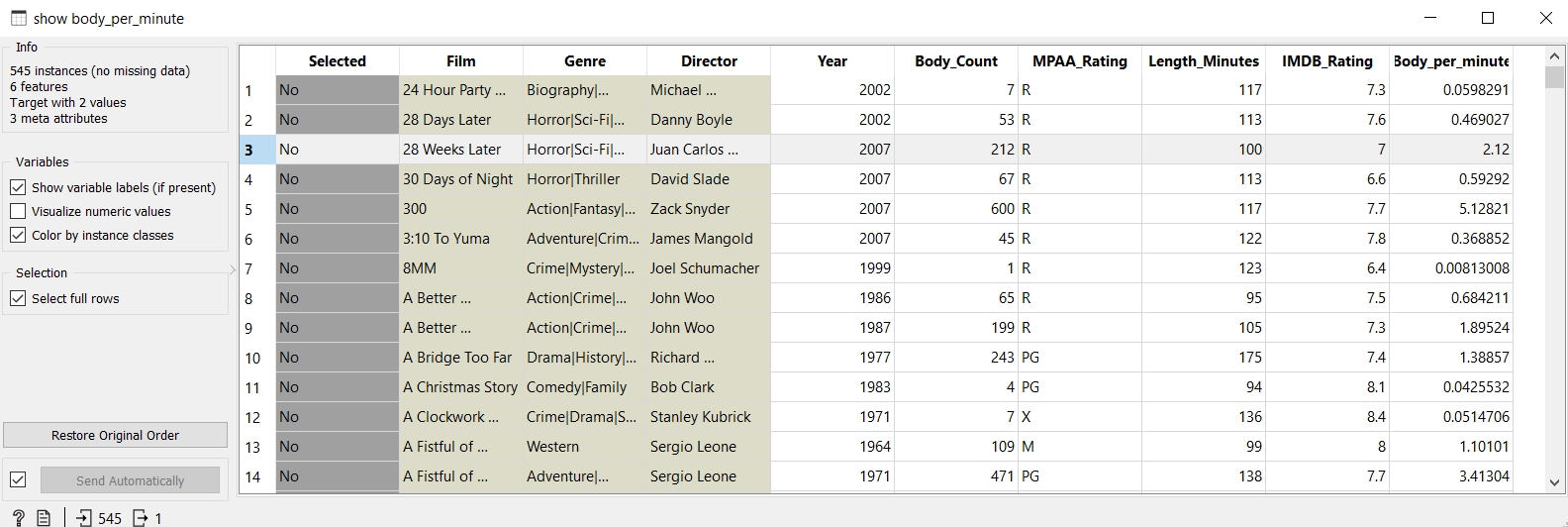
**Досліджуваний датасет:**

****

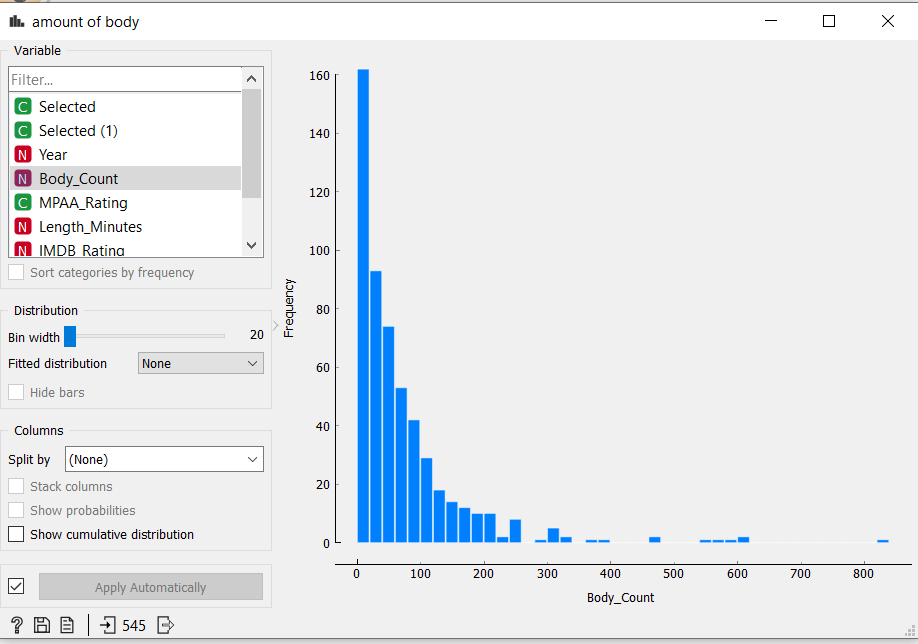
Щоб визначити кількість смертей за хвилину, потрібно вставити формулу розрахунку



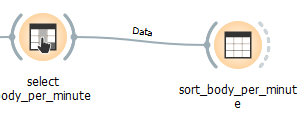
Після чого отримуємо табличку з новою колонкою смертей за хвилину для кожного фільму:

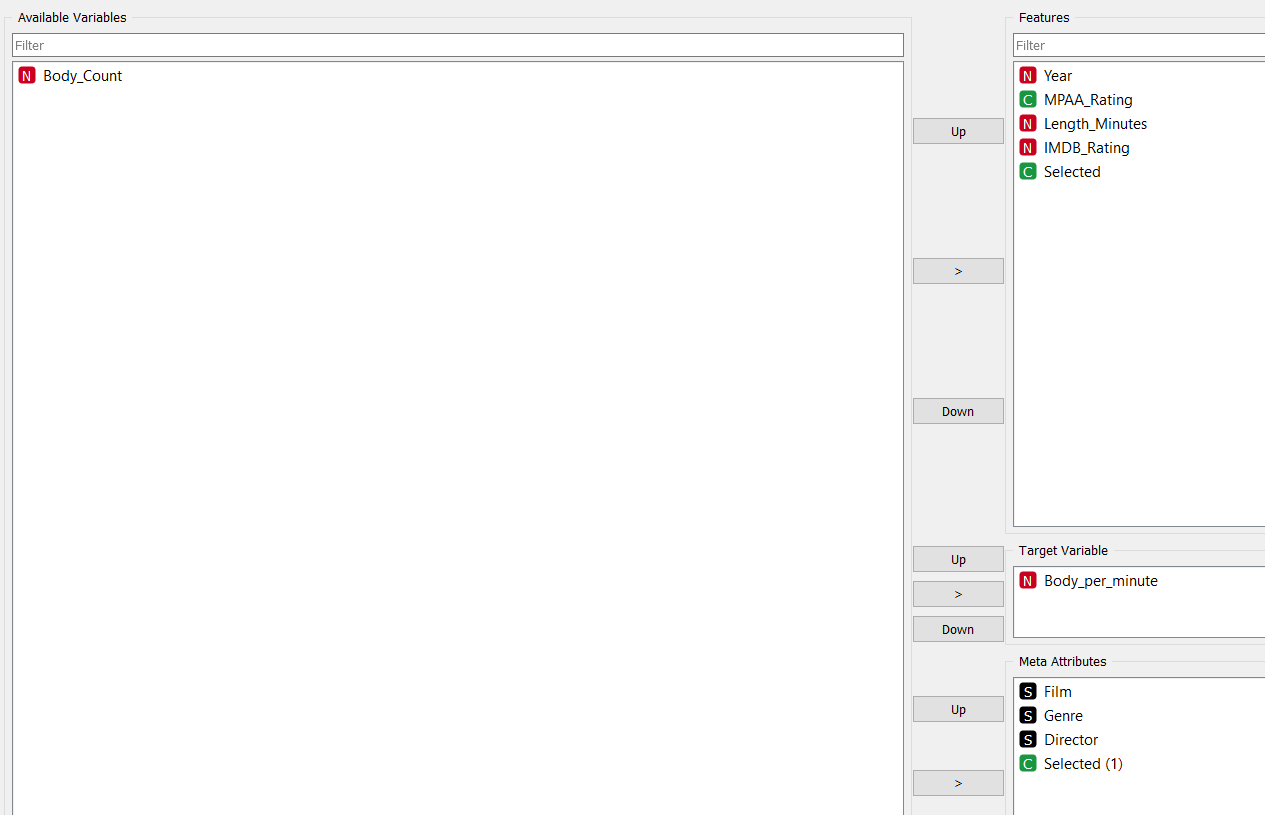


Гістограма частоти кількості смертей у фільмі



Вибираємо колонку смертей за хвилину як “primary”

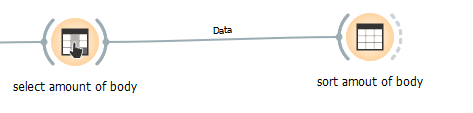


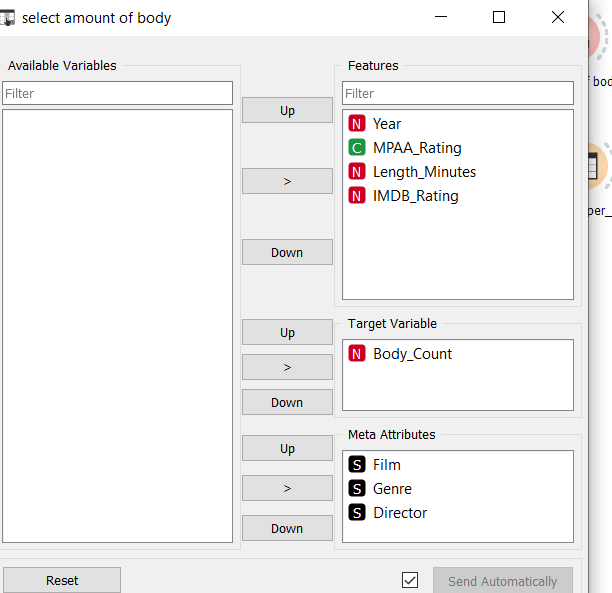


І отримаємо такий датасет:

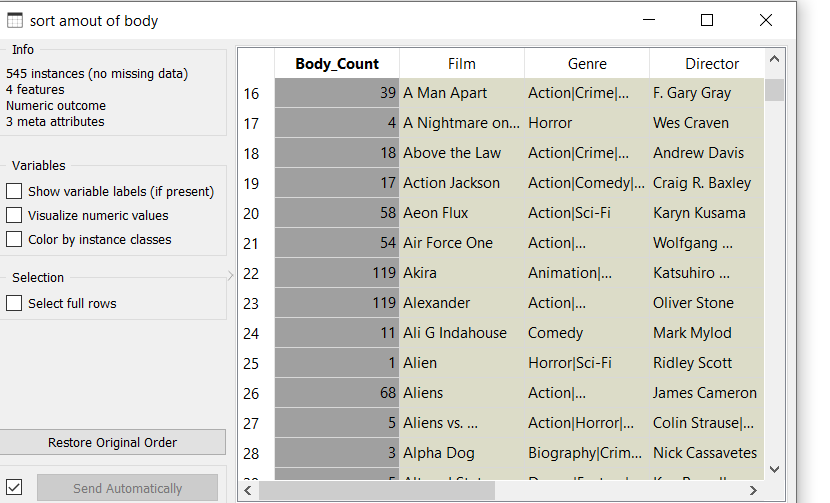


Тапер зробимо те саме, тільки по загальній кількості смертей у фільмі:

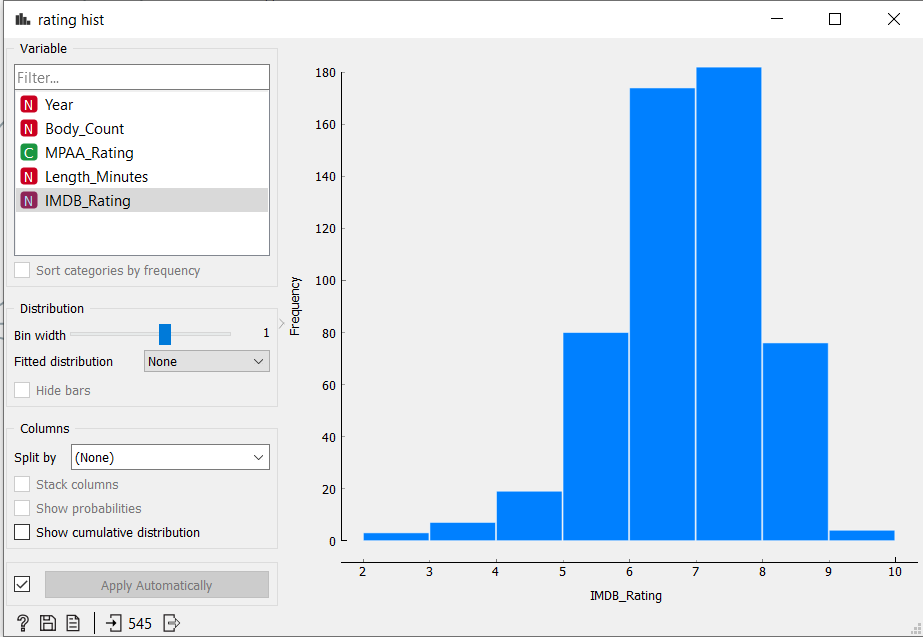




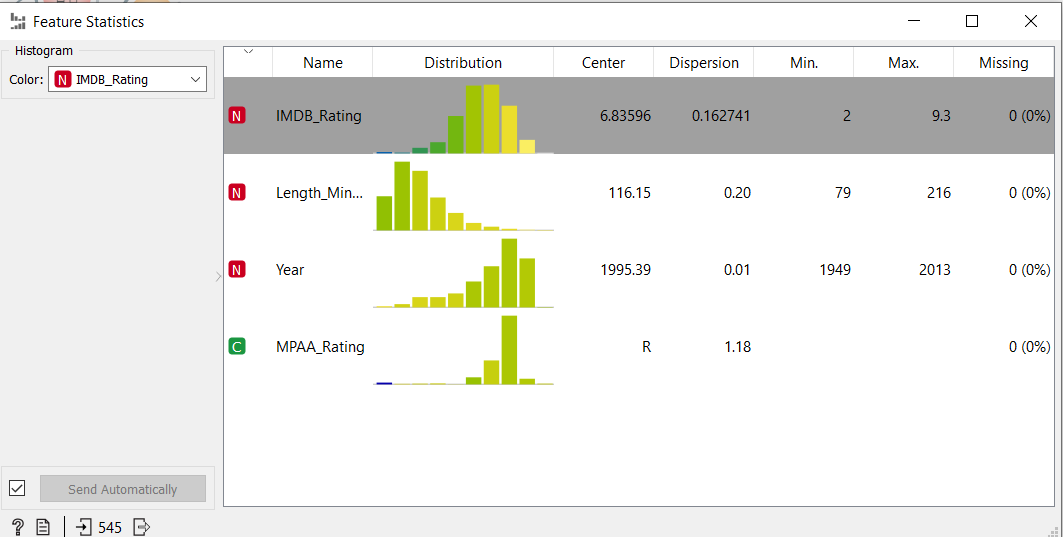
І отримаємо:



Побудуємо гістограму IMDB рейтингу:

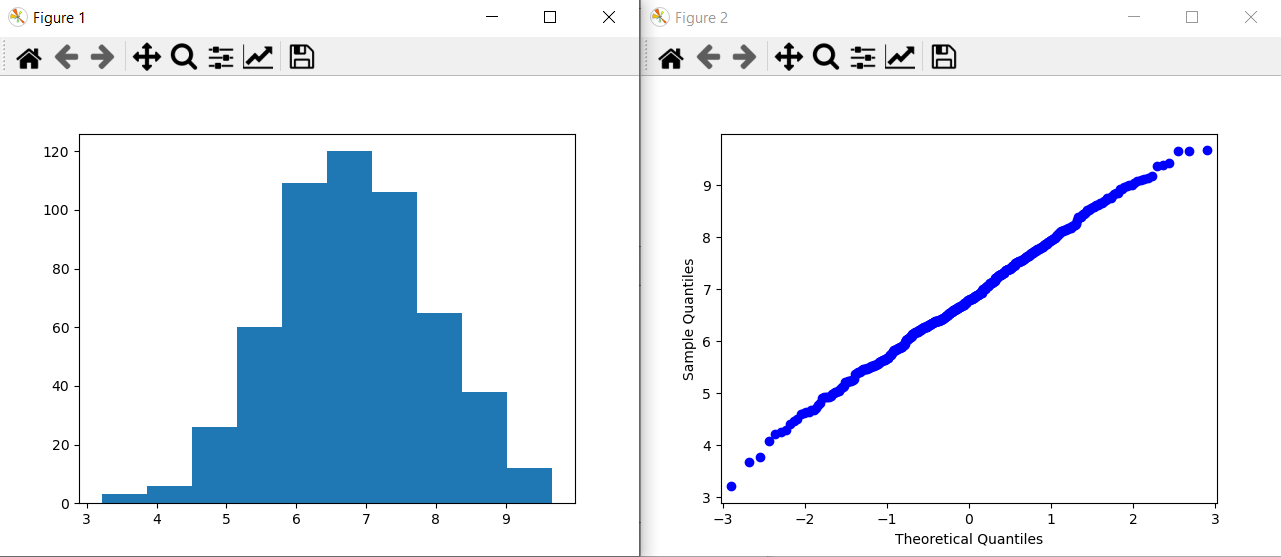


Виведемо деякі статистичні дані для колонок для яких ми встановили фільтри:

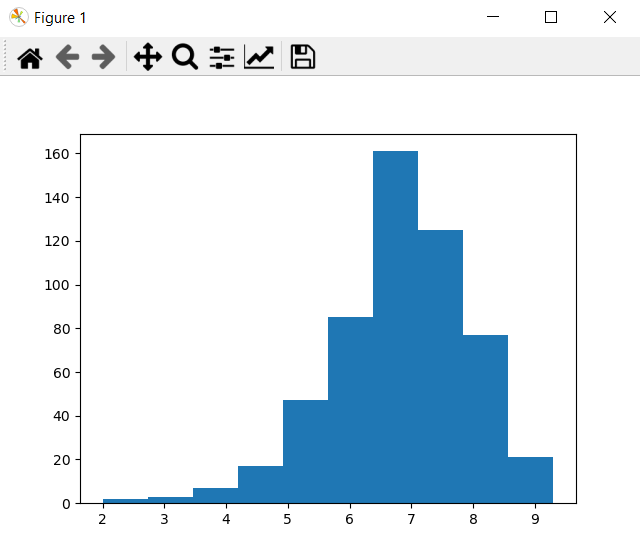


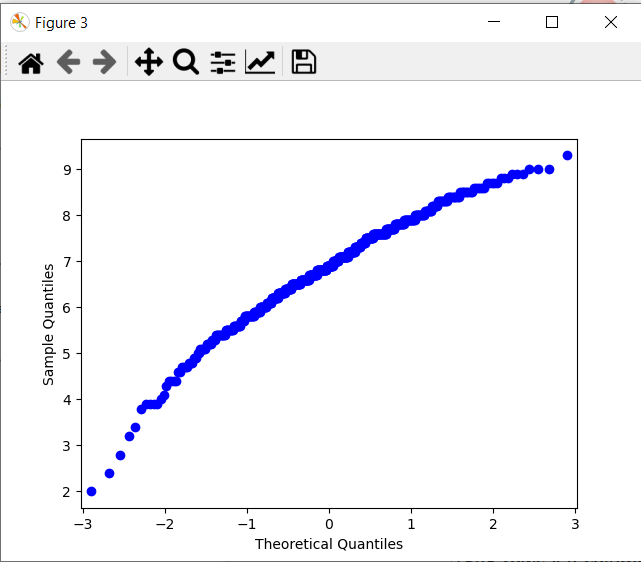
Тепер за допомогою Python скрипта візуалізуємо розподіл для симульованої та для нашої вибірки (IMDB рейтинг).

Симуляція:



Дана вибірка:





Порівнявши графіки розподілу, різниця у нормальності очевидна.

Код для реальних даних:

import random

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import statsmodels.api as sm

df1 = pd.DataFrame(data)

imdb\_mean = df1[3].mean()

imdb\_std = df1[3].std()

random.seed(900)

imdb\_simulation = np.random.normal(imdb\_mean, imdb\_std, df1.shape[0])

df = pd.DataFrame(imdb\_simulation)

plt.hist(df[0], bins=10)

plt.show()

sm.qqplot(df[0])

plt.show()

Код для симуляції нормального розпроділу

import matplotlib.pyplot as plt

import statsmodels.api as sm

import pandas as pd

data = in\_data.copy()

df = pd.DataFrame(data)

print(df)

sm.qqplot(df[3])

plt.show()

Висновок: як на мене, Orange - це потужний інструмент для аналізу та візуалізації даних. Головна його перевага – це простота. Для того, щоб виконати це завдання з допомогою лише Python, тут знадобляться базові знання програмування та мови, чого не потрібно при використанні Orange.