МІНІСТЕРСТВО ОСВІТИ І НАУКИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кафедра систем штучного інтелекту



Лабораторна робота №5

З курсу “Візуалізація даних”

Виконав:  
студент групи КН-310

Бурак Марко

Викладач:

Виклюк Я.І.

Львів – 2020

**Тема:** Згруповані розподіли частот.

**Мета:** Побудова розподілу частот за емпіричними даними.

**Завдання:** Представити в табличній та графічній формі диференціальний та інтегральний розподіли частот за табличними даними:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 115 | 109 | 119 | 96 | 114 | 91 | 92 | 83 | 83 | 128 | 100 | 107 | 90 | 88 | 105 | 93 | 105 | 103 | 97 | 106 |
| 112 | 104 | 116 | 85 | 106 | 89 | 102 | 95 | 102 | 92 | 112 | 89 | 78 | 83 | 92 | 77 | 97 | 120 | 114 | 89 |
| 97 | 80 | 99 | 86 | 96 | 112 | 102 | 117 | 90 | 99 | 80 | 80 | 83 | 87 | 93 | 84 | 87 | 79 | 96 | 114 |

Пояснення: кожен студент повинен до кожної цифри у клітинках добавити останню цифру порядкового номера у журналі, але з певними корективами: 307 – (+1), 308 – (+2), 309 – (-1), 310 – (-2).

Мій порядковий номер в журналі – 4, остання цифра – 4, група – КН-310.

Код програми:

**Імпортую бібліотеки для роботи з датасетом**

import pandas as pd

import matplotlib.pyplot as plt

import math

import numpy as np

**Додаю до кожного значення з датасету 2, як сказано за умовою задачі**

correct\_number = 4 - 2

**Функція програми**

def analys(data):

# Data analys

n = len(data.index)

min = data["Values"].min()

max = data["Values"].max()

k = math.ceil(1 + 3.32 \* math.log10(n))

l = (max - min) / k

print("Volume of data:", n)

print("Min value:", min)

print("Max value:", max)

print("Amount of classes:", k)

print("Interval:", round(l, 3))

data\_table = (data.groupby(pd.cut(data["Values"], np.linspace(min-0.01, max, num=k+1))).count())

summa = data\_table["Values"].sum()

data\_table["Values\_%"] = (data\_table["Values"] / summa \* 100)

data\_table["Cumulative\_Sum"] = data\_table["Values"].cumsum()

data\_table["Cumulative\_Sum\_%"] = data\_table["Values\_%"].cumsum()

print(data\_table)

**У цій функції я формую датасет з даних Values, Values%,Cumulative\_Sum та Cumulative\_Sum\_%**

Values відповідає за кількість значень у інтервалі, значення у датасеті це бали IQ.Values%, відсоткове представлдення Amount.Cumulative\_Sum та Cumulative\_Sum\_% це кумулятивна, або ж накопичувальна сума та відсодкове представлення цієї суми , відповідно.

Також я порахував обсяг вибірки, що рівний 60, це означає що наш датасет вміщає 60 різних значень.

Підрахував мінімальне та максимальне значення з вибірок, 77 і 128 відповідно

Опісля підрухував кількість класів за допомогою формули Стерджеса K = 1+3.32\*lg n, де n - це обсяг вибірки.

Ну і визначив інтервали датасету, які теж рахуються за формулою ℷ = (max – min )/ k ,де к це кількість класів.

Кількість класів вийшла 7, а інтервал 7.286

**Ця частина коду, відповідає за візуальне зображення графіків, після апробації, програма зображає 4 гістограми а також один графік який містить у собі гістограму та пряму.**

fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(10, 10))

data\_table.plot(y="Values",ax=axes[0, 0], kind="bar", title="Amount of rating")

plt.xticks(rotation='90')

data\_table.plot(y="Values\_%", ax=axes[0, 1], kind="bar", title="Amount\_% of rating")

plt.xticks(rotation='90')

data\_table.plot(y="Cumulative\_Sum", ax=axes[1, 0], kind="bar", title="Cumulative\_Sum of rating")

plt.xticks(rotation='90')

data\_table.plot(y="Cumulative\_Sum\_%",ax=axes[1, 1], kind="bar", title="Cumulative\_Sum\_% of rating")

plt.xticks(rotation='90')

plt.show()

me = data\_table.plot(y="Cumulative\_Sum\_%", kind="line")

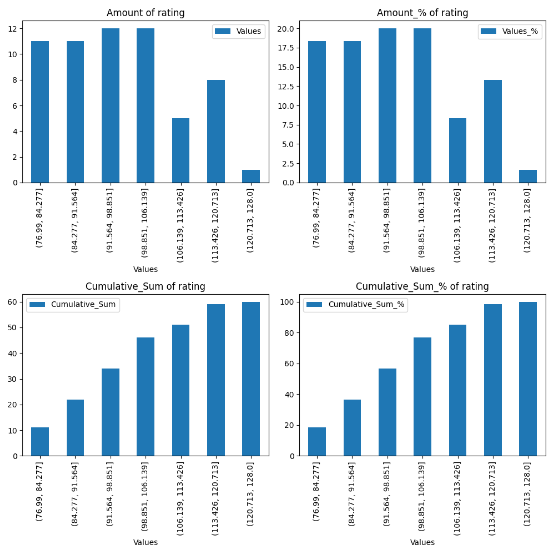
data\_table.plot(y="Values\_%", kind="bar",ax=me)

me.grid()

plt.show()

analys(pd.read\_csv("data.csv"))

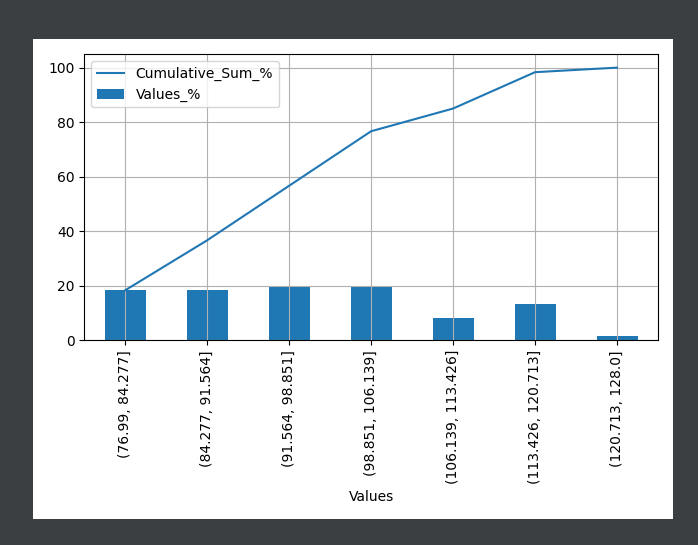
**Після апробації програми отримаємо такі графіки:**



З першого графіка можна побачити, що здебільшого у вибірці знаходяться значення від 91.56 до 106.14. Та бачимо, що після 106, відбувається різкий спад до 113.4, проте росте знову з 113.4 до 120.7

Найменше ж людей склало IQ тест з 120.7 до 128, так як це максимальне значення з вибірки.

Також з нижніх графіків видно на скільки саме відбувається зріст між сусідніми значеннями.

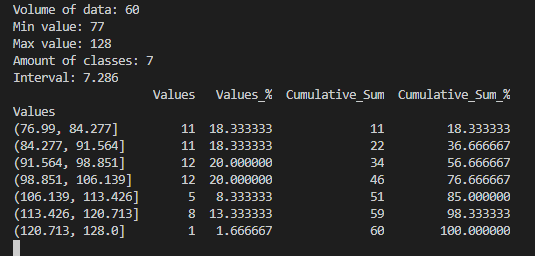


Графік деференціального та інтегрального розподілу частот

Деференціальні розподіли представляють значення частот окремо для кожного х.  
Інтегральні (накопичувальні, кумулятивні) це додавання усіх попередніх х частот.

Площо, яка знаходиться під диференціальним розподілом має сенс частоти.

З графіка вище можна побачити, що значення з 77 до 84.27 включно становлять приблизно 18% від усієї вибірки, з 77 до 91.564 включно становлять приблизно 37% вибірки, а від 77 до 106.139 - 77%, і від 77 до 128 включно, 100%, тобто уся вибірка. Цей графік дуже зручний, щоб проводити таку аналітику



Тут представлена таблиця точних відсоткових значень для кожного інтервалу, можна зробити висновок, що Графік деференціального та інтегрального розподілу частот дійсно зручний для аналітики, адже ми отримали близькі до реальних значень числа.

**Висновок:** Я навчився будувати та візуалізувати розподіл частот за емпіричними даними, попередньо згрупувавши дані, за допомогою бібліотеки Pandas мовою програмування Python.