UDC 004

Mykola Turchyn, Master's Student,
Olena Chyzhmotria, Senior Lecturer,
Iryna Dmytrenko, Senior Lecturer
Zhytomyr Polytechnic State University

SEABORN AS A TOOL FOR EFFECTIVE WORK WITH CATEGORICAL DATA

In the fields of data science and machine learning, visualization is a key tool for identifying patterns and presenting results. This is especially true when working with categorical data, which represents discrete groups. The Seaborn library, based on Matplotlib, has become one of the most popular solutions for creating such graphics. It simplifies the creation of aesthetic statistical diagrams, integrates closely with Pandas DataFrames, and provides high-level functions that reduce code volume compared to Matplotlib [1].

Categorical Data

Categorical data represents discrete groups or categories, such as gender, product types, or geographic regions. Adequate visualization of categorical data allows us to identify distributions, compare values across categories, and draw meaningful insights. Bar plots, box plots, and violin plots are the most effective tools for this purpose [2]. An example is shown in Fig. 1.

Bar plots are a very popular way for visualizing categorical data, where one axis represents the categories, and the other shows the corresponding numerical values. They make it easy to compare values across groups and identify trends or differences. On the other hand, count plots display the frequency of each category, which is particularly useful for interpretation of the distribution and prevalence of different groups.

Another way to examine the distribution of the numerical values across categories is through the use of box plots and violin plots. Box plots summarize data using the median, the quartiles, and a limit (or indicator) of the outliers, thus permitting the observation of variation or deviations from the data. In a similar line, box plots supplemented with kernel density estimates called violin plots which provide information on several cases and their distribution across categories. Seaborn is a very relevant tool for the interpretation of trends and variations at a group level, thanks to its ability to provide detailed and versatile analysis of categorical data.

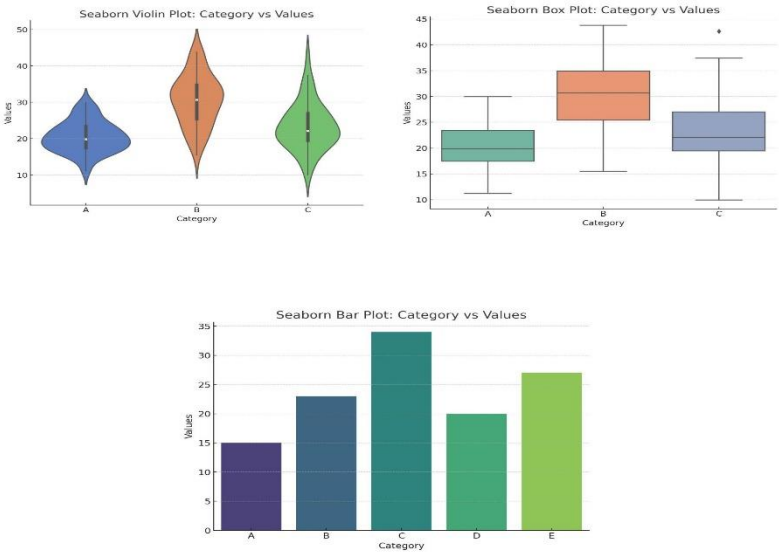

Fig. 1 – Example of visualization of categorical data

Summary

Seaborn can be useful for visualizing and analyzing categorical data. It features friendly syntax, integrates well with Pandas, and has good high-level plotting functions, making things easier when creating informative visualizations. If you are interested in studying data distributions or comparing values across categories, Seaborn has the tools to expose and present interesting findings. By learning the essential functions for categorical data, such as bar plots, box plots, and violin plots, you will be able to convert unorganized data into clear and compelling messages that aid in making decisions.

References:

1. Wes M. Python for Data Analysis Data Wrangling with pandas, NumPy, and Jupyter / McKinney Wes., 2022. – 561 c.

2. A Beginner's Guide to Seaborn for Data Visualization in Python [Information resource] / DataCamp – Resource access mode: https://www.datacamp.com/tutorial/seaborn-python-tutorial.