

# Explainable Similarity Correlation of Categorical Data and Bar Charts

Imre Rudas<sup>1</sup> and Ildar Batyrshin<sup>2</sup>

<sup>1</sup>Obuda University, Budapest, Hungary

<sup>2</sup>Instituto Politecnico Nacional, Mexico

## Abstract

Social and behavioral sciences, business analytics, computational linguistics, machine learning, and pattern recognition widely use categorical data. Such data are often characterized by distributions of counts or frequencies of appearance of variable measurements in different categories or classes. An analysis of possible relationships between such categorical data distributions is essential for many applications. For example, one can compare the distributions of votes for several political parties in two states or regions, the distributions of gender preferences for social or professional activity, distributions of preferences for two types of customers, distributions of sales of car models in two countries, etc. Usually, such distributions are presented by contingency tables or bar charts. Descriptive statistics uses bar charts for visual comparison and analysis of distributions. Statistics has several measures for analyzing the possible association between categorical variables. In our work, we propose the method of calculating the correlation between categorical data and bar charts. The proposed approach is based on recently developed methods of construction of correlation functions from fuzzy similarity and dissimilarity functions. We propose a new approach to the presentation of bar charts visually explaining the detected correlation between them. We hope that the proposed methods of analysis of relationships between categorical data will find wide applications in business analytics and other areas where categorical data appear.