組別 Team ID：202221

專題屬性 Category：管理類（Management）

專題名稱 Project：Data Visualization for Dimension Reduction

一、指導老師 Advisor：洪朝貴老師（Prof. Chao-Kui, Hung）

二、 組員 Team members：

巴凱莉（10814180）、鍾翰霖（10814207）、滿都海（10814245）

三、 系統環境 System environment：

（一）軟體 Software：

作業系統 Operating System： Windows, Linux

語言 Programming language： Python

開發工具 Tool kits： Umap, Pandas, Numpy, Matplotlib, Seaborn

（二）硬體 Hardware：

Computer

四、 簡介：

（一）系統簡述

現在的問題是大多數數據集都有大量的變量。換句話說，它們具有大量數據分佈的維度。數據可視化可能變得具有挑戰性，並且通常幾乎不可能手動完成。然而，這種視覺探索在任何與數據相關的問題中都非常重要。因此，了解如何可視化高維數據集是關鍵。這可以使用稱為降維的技術來實現。在這個項目中，我們的目標是使用 UMAP（統一流形近似和投影）為幾個大數據集生成有意義的可視化。 UMAP 是一種非線性降維方法，對於可視化集群或數據點組及其相對接近度特別有用。對於這個項目的起點，我們必須選擇具有大量數字和分類列以及數千行的合適數據集。進一步的過程包括預處理這些數據集並將它們輸入環境、分配顏色、使可視化具有交互性等。通過研究數據點是如何組織和分組的，我們從可視化的結果中得出了幾個重要的假設。

（二）特色

- 互動情節。
- 優化的代碼行。
- 參數更改簡單。

五、 Introduction：

Introduction

The problem today is that most data sets have a large number of variables. In other words, they have a high number of dimensions along which the data is distributed. Data visualization can then become challenging and is often nearly impossible to do manually. However, such visual exploration is incredibly important in any data-related problem. Therefore, it is key to understand how to visualize high-dimensional data sets. This can be achieved using techniques known as dimensionality reduction. In this project, we aim to employ UMAP (Uniform Manifold Approximation and Projection) to generate meaningful visualizations for several big datasets. UMAP is a nonlinear dimensionality reduction method that is particularly useful for visualizing clusters or groups of data points along with their relative proximities. For this project's starting point, we had to select suitable datasets with a large number of numerical and categorical columns and thousands of rows. Further processes include preprocessing those datasets and feeding them into the environment, assigning colors, making visualization interactive, etc. By studying how the data points have been organized and grouped, we drew several significant assumptions from the visualization's outcome.

Features

- Interactive plot
- Optimized lines of code
- Easy parameter changing