COMS4060A/7056A:

University of the Witwatersrand 2023

October 15, 2023

Due: November 6, 2023

1 Introduction

This assignment is based on content covering dimensionality reduction. You will be required to perform some basic data cleaning and exploration techniques on prescribed datasets.

There is no strict requirement on the format of your submission, but any answers should be reasoned, discussed, and relevant data or results should be provided to substantiate this. You might like to submit either a PDF or a Jupyter notebook, for example.

You can use any programming language or tool you would like, however Submissions should include both text and code. You may use any language that you like.

Key Features:

1. Player Information: Player name, team(s) played for during the season.
2. Per Game Statistics: A wide array of per-game statistics, including points scored (PPG), assists (APG), rebounds (RPG), steals (SPG), blocks (BPG), and more.
3. Shooting Efficiency: Metrics like field goal percentage (FG%), three-point percentage (3P%), two-point percentage (2P%), and free throw percentage (FT%) for assessing scoring efficiency.
4. Advanced Statistics: A wide array of advanced metrics such as value over replacement player (VORP), win shares (WS) and true shooting percentage (TS/
5. Salaries: The financial aspect of the dataset includes player salaries for the 2022-23 season, offering insights into player earning

2 Instructions

This is a kaggle dataset of NBA player salaries.

1. Do a preliminary analysis of the dataset. Are any points which seem odd. Decide weather or not to remove them and justify your decision.
2. Discuss various methods of handling the missing values, with a view towards plotting the data in a two or three dimensional plot.
3. Reduce the data to three dimensions in **three** of the following different ways.
   (a) tSNE
   (b) UMAP
   (c) Using an autoencoder
   (d) MDS (and variants)
   (e) ISOMAP
   (f) LLE.
4. For each method plot the dataset in the reduced space.
5. Comment on the similarities and differences in these plots.
6. Which method do you think worked best? Explain why.