Summary of *Veridical Data Science*, Chapters 5–6 [*]

Shizhe Zhang (3041882158)
shizhe_zhang@berkeley.edu

February 5, 2026

Chapter 5: Exploratory Data Analysis

EDA is a question-and-answer process where you use visualizations to explore the data and generate hypotheses. Visualizations help you spot patterns, trends, outliers, and relationships that are hard to see from tables or summary statistics alone. And there could be many different ways to visualize the same data, each highlighting different aspects. Choosing the one that best delivers the message is a key part of EDA.

Poor choices of the visualization, like inconsistent scales, would hinder interpretation. Judgment calls also come into play in EDA, such as deciding how to bin data for a histogram or how to handle outliers. And we should test the stability of our judgment calls by trying different visualizations and checking if the patterns hold up.

Chapter 6: PCA

PCA is a useful tool when dealing with high-dimensional data. PCA creates new variables, called principal components, which are linear combinations of the original variables and capture as greatest variation in the data as possible. Note that the variable created by PCA is not necessarily interpretable, and we forfeit some information when we only keep the top few components, so we need to be careful when using PCA for dimensionality reduction.

There are also judgment calls in PCA, such as mean centering and SD scaling, which can dramatically change PCA results. Whether to perform these steps depends on the context and the goals. We can still apply the PCS framework to PCA: we should check if the patterns revealed by PCA are stable under different preprocessing choices, and we should interpret the results carefully, keeping in mind that PCA is an exploratory tool that reveals structure but does not automatically explain causes.

[*] `https://vdsbook.com/`