* [Overview](#gjdgxs)
* [Preparation Reading](#30j0zll)

# CSE 450 - Machine Learning & Data Mining

## Module 02 — Data Preprocessing and Clustering

 [*Photo by Edgar Chaparro on Unsplash*](https://unsplash.com/photos/r6mBXuHnxBk)

## Overview

Estimated Reading Time

Plan on around 90 - 120 minutes for this preparation reading, which consists of a mix of online and textbook reading.

The objective of this module is to provide a real-world scenario in which you can practice the following data science / machine learning skills:

* Assessing Data Quality
* Investigating Feature Relationships
* Scaling and Normalization
* Ethics
* Unsupervised Learning and K-Means Clustering

## Preparation Reading

Complete the following preparation reading:

* Read sections 3.1, 3.2, 3.3, 3.4, and 3.6 of your text. (You've already read section 3.5, but feel free to review it.)
* Watch the following introduction to Cluster Analysis:
* Read about [the Sci-Kit Learn implementation of K-Means clustering](https://scikit-learn.org/stable/modules/clustering.html#k-means).
* Review [the Sci-Kit Learn API docs for K-Means clustering](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans)
* Read this article: [Top three mistakes with K-Means Clustering during data analysis](https://zerowithdot.com/mistakes-with-k-means-clustering/)
* Read this article: [K-means clustering is not a free lunch](http://varianceexplained.org/r/kmeans-free-lunch/)