DATA SCIENCE AND MACHINE LEARNING IN MANAGER SELECTION AND PORTFOLIO CONSTRUCTION

Cordell L. Tanny, CFA, FRM, FDP
October 19, 2021

MY OBJECTIVES

1. Demonstrate that not every data science or machine learning (ML) project must be a massive endeavour.
2. Show how powerful Python is for data analysis and visualization (exploratory data analysis, or EDA).
3. Demonstrate how a hierarchical clustering algorithm can be used to help with manager due diligence and portfolio construction.

STEP 1: IDENTIFYING THE BUSINESS NEED

The first step in any data science project is to identify the business need:

What are you trying to accomplish, and do you have the data?

WHAT AM I TRYING TO SOLVE?

- Can I combine highly correlated managers in a way that increases portfolio diversification?
- Specifically, identifying managers who try to beat the same benchmark by using different strategies.

Some context

- Portfolio construction relies heavily on three inputs:
  - Returns
  - Risks
  - Correlations
- Correlations are unstable, subject to estimation error, and don’t describe potential relationships.

STEP 2: EXPLORATORY DATA ANALYSIS

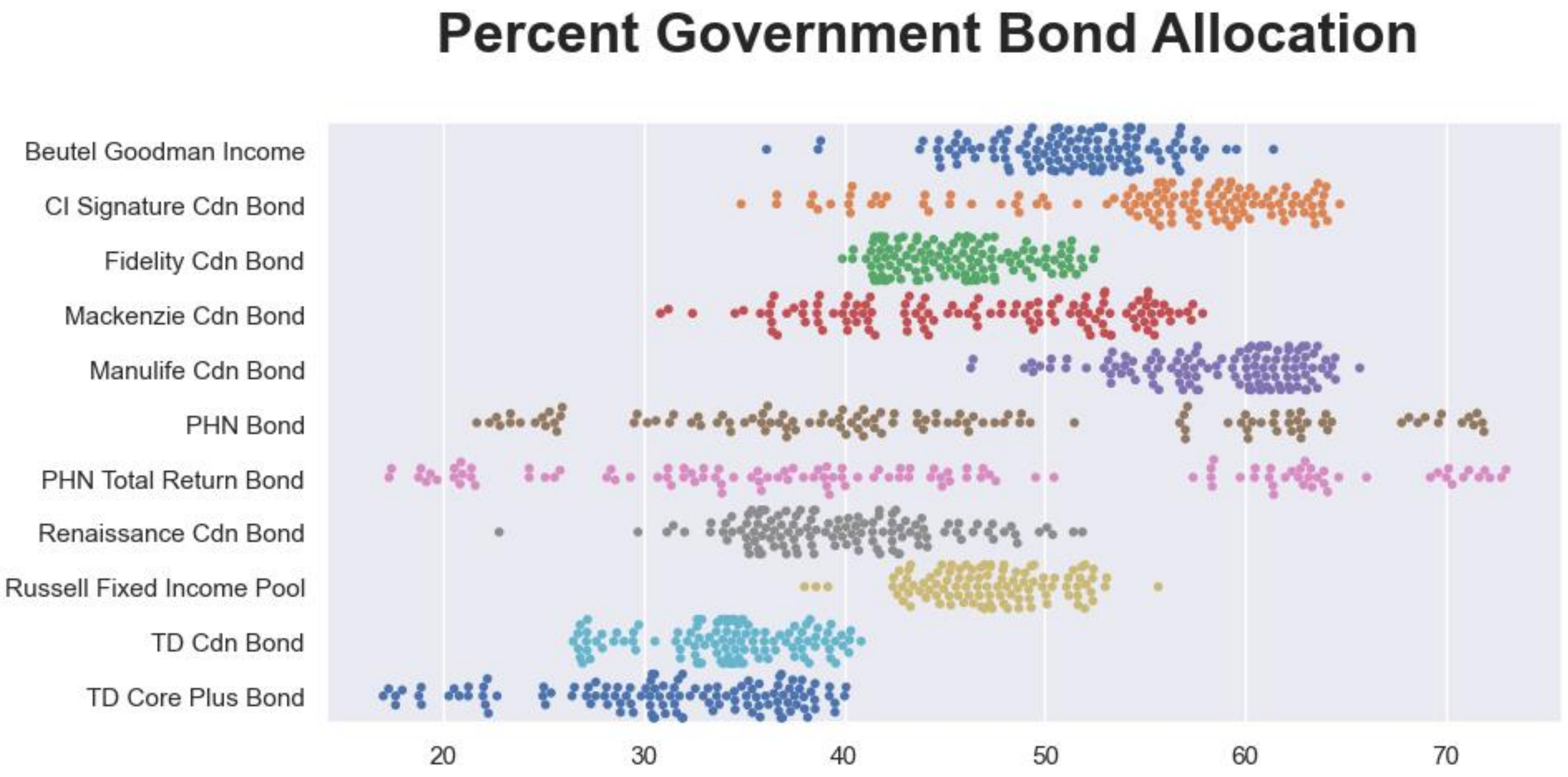
Understand your data set:

Explore it visually to help determine the underlying relationships, dimensions, and shortcomings.

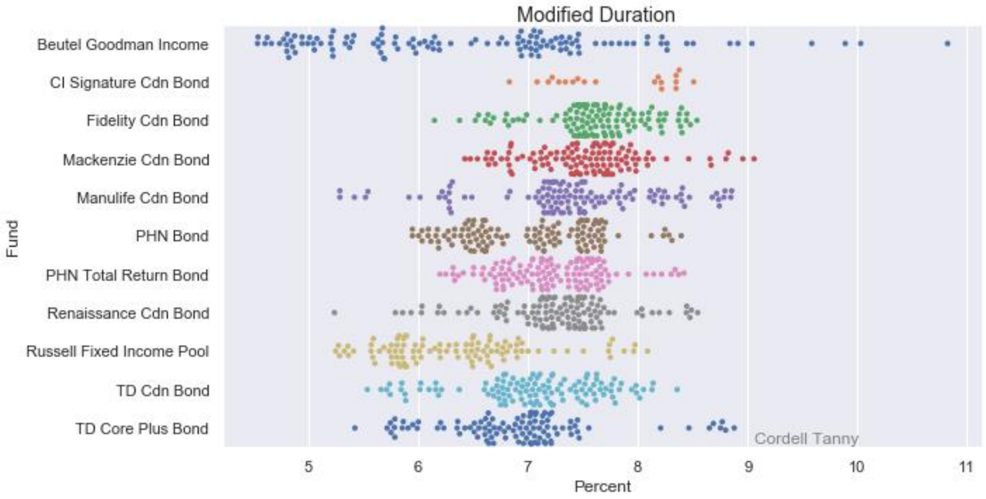
MANAGER CHARACTERISTICS OVER TIME

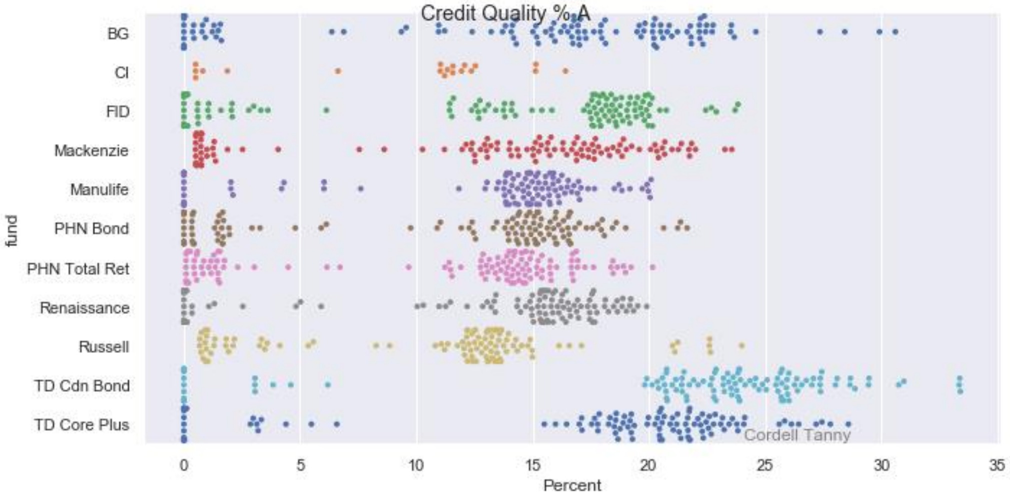
- EDA can help us identify and confirm the manager’s strategy.
  - For example, identifying a duration manager or a sector allocator.
- Retrieve time series data for all the features that you wish to analyze.
- Potential problems:
  - Missing data
  - Different reporting frequencies
  - Insufficient history
- Python and R have many built-in functions to help with data cleaning and preparation.

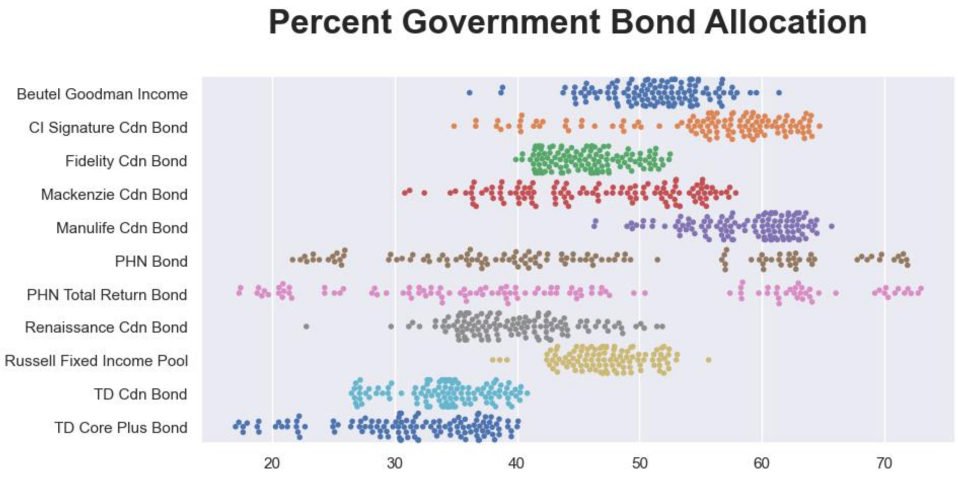
FEATURE ANALYSIS



WITH MULTIPLE FEATURES







OUR FEATURES

% Government	% Corporate	% Cash	% Securitized	% Canada
% US	% Emerging Markets	Modified Duration	# of Bond Holdings	Average Coupon
% AAA	% AA	% A	% BBB	% BB
% B	Yield to Maturity			

STEP 3: MACHINE LEARNING

Finding similarities across 15 dimensions and 60 funds is a perfect task for a machine learning algorithm.

But which one?

SUPERVISED VS. UNSUPERVISED LEARNING

Supervised Learning

- We have x variables (features) and a y variable (target or label).
  - Ex: linear regression, logistic regression.
- Used for predictions, inferences, understanding relationships between features.
- Categorical or continuous variables.
- Random forests, gradient boosting, ANN.

Unsupervised Learning

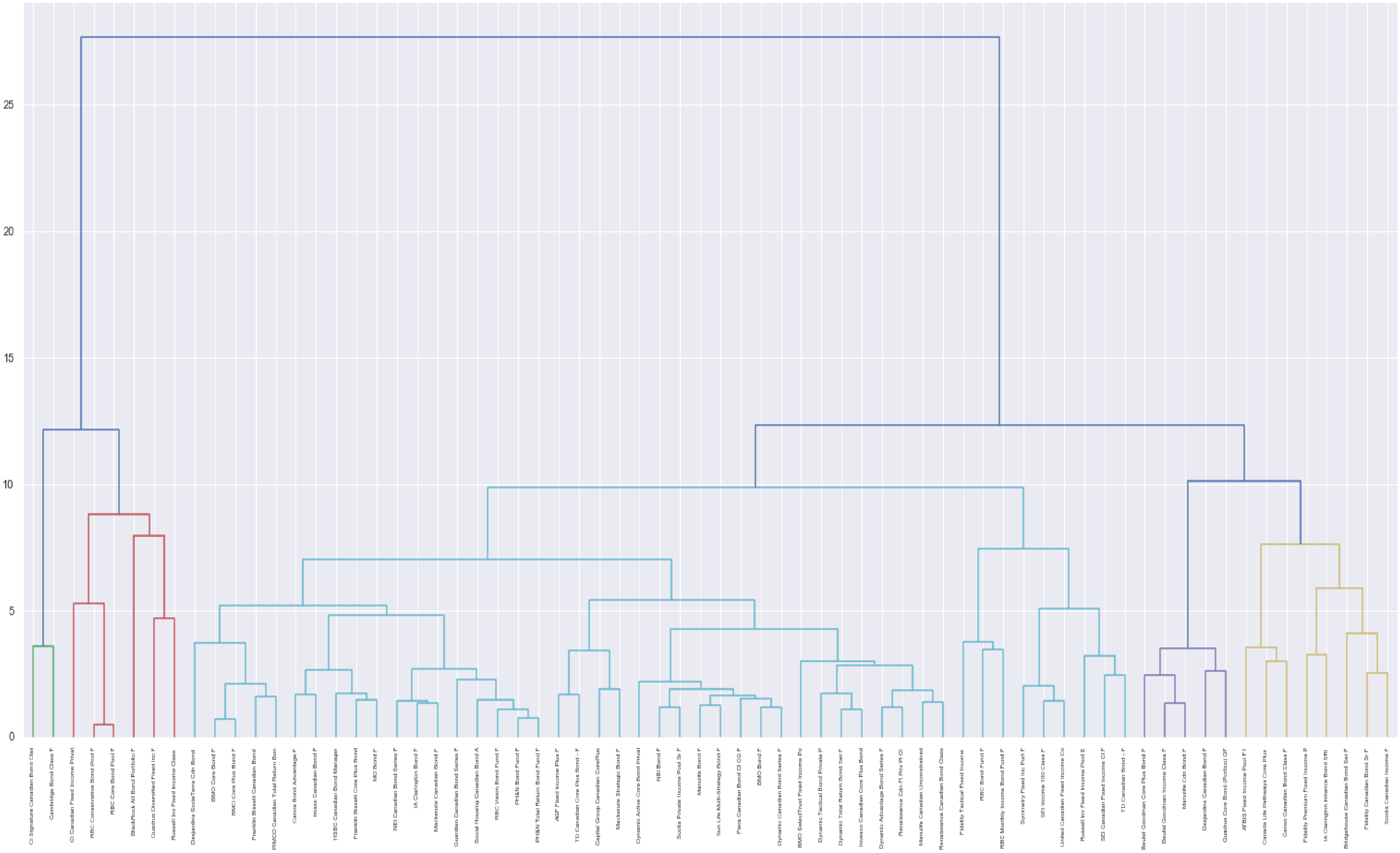
- We have features, but no targets.
- Used for identifying patterns within unlabeled data.
- Can be used for cluster analysis.
- Hierarchical clustering, k-means cluster analysis.

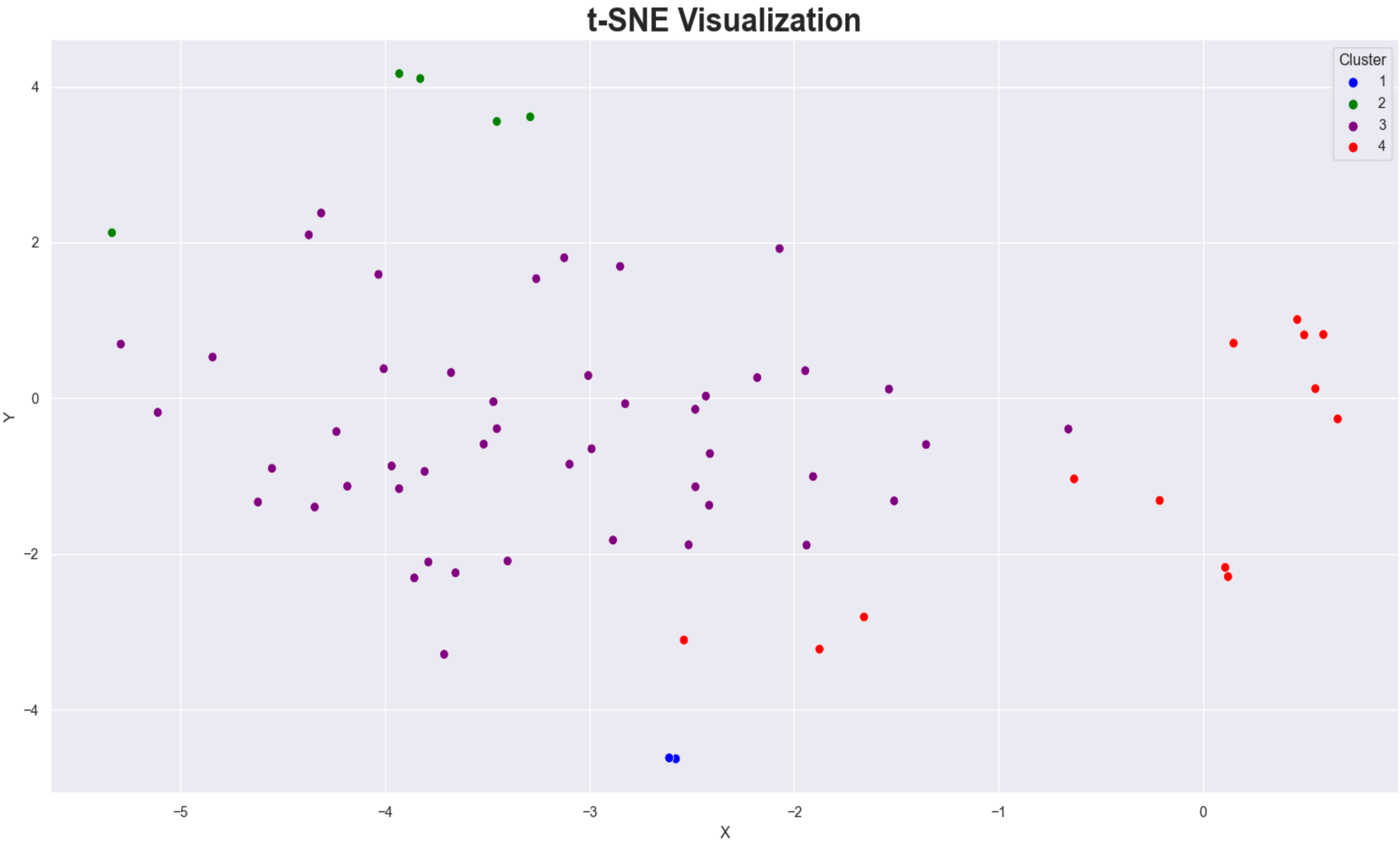
HIERARCHICAL CLUSTERING

Recap:

- We have unlabeled data consisting of 15 features over a 12-year period for over 50 funds.
- An unsupervised learning method should be used since we are trying to uncover the relationships between the funds based on how these features have changed over time.
- Enter hierarchical clustering:
  - The algorithm groups similar observations (in this case, funds) into clusters.
  - We get fewer clusters with more observations as we move up the hierarchy.

Canadian Fixed Income Category Dendrogram




t-SNE Visualization
Cluster
● 1
● 2
● 3
● 4
4
2
0
−2
−4
Y
−5
−4
−3
−2
−1
0
X

WHAT HAVE WE ACCOMPLISHED?

1. Uncovered the substructure within our investment category.
2. Used the dendrogram/t-SNE to find managers in different clusters that could offer diversification benefits.
3. Discovered a method to complement optimization where we can add on to the concept of correlation.
4. Used EDA to visualize our features in different kinds of graphs that would be near impossible or time consuming to do in excel.

MY OBJECTIVES

1. Demonstrate that not every data science or ML project has to be a major project.
2. Show how powerful python is for data analysis and visualization (exploratory data analysis, or EDA).
3. Demonstrate how a hierarchical clustering algorithm can be used to help with manager due diligence and portfolio construction.