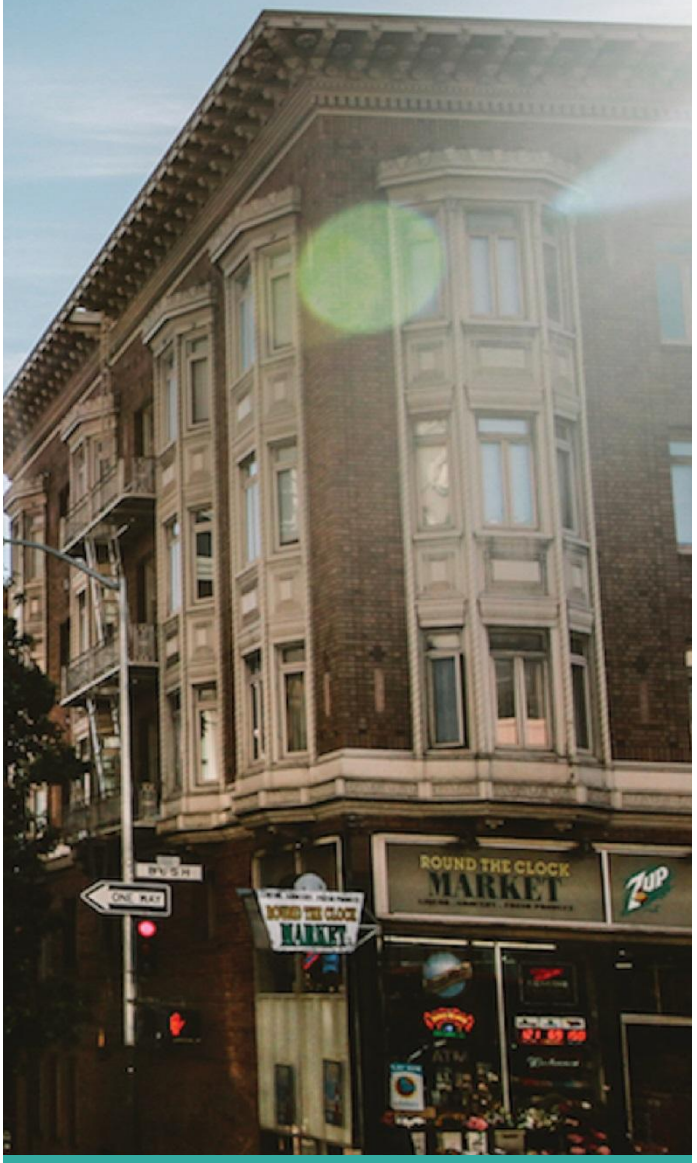
LAB 4: CLUSTERING



OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

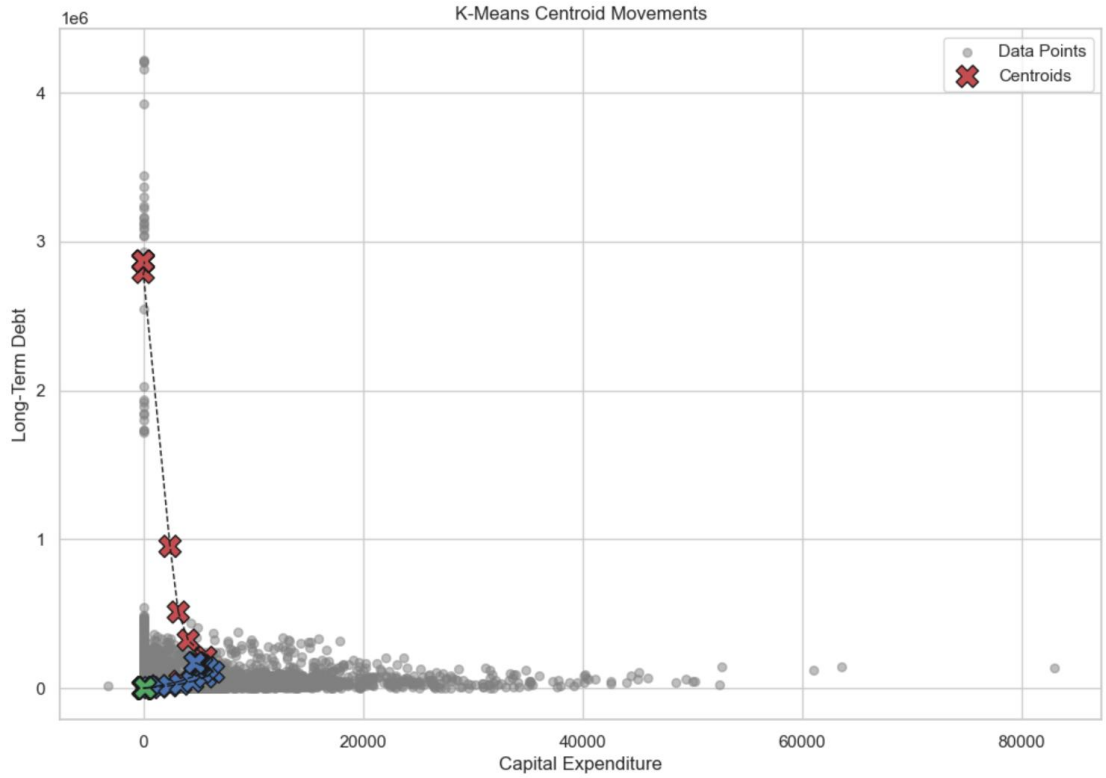
Authored by: Andres Sebastian Gaibor Heredia

1. Clustering

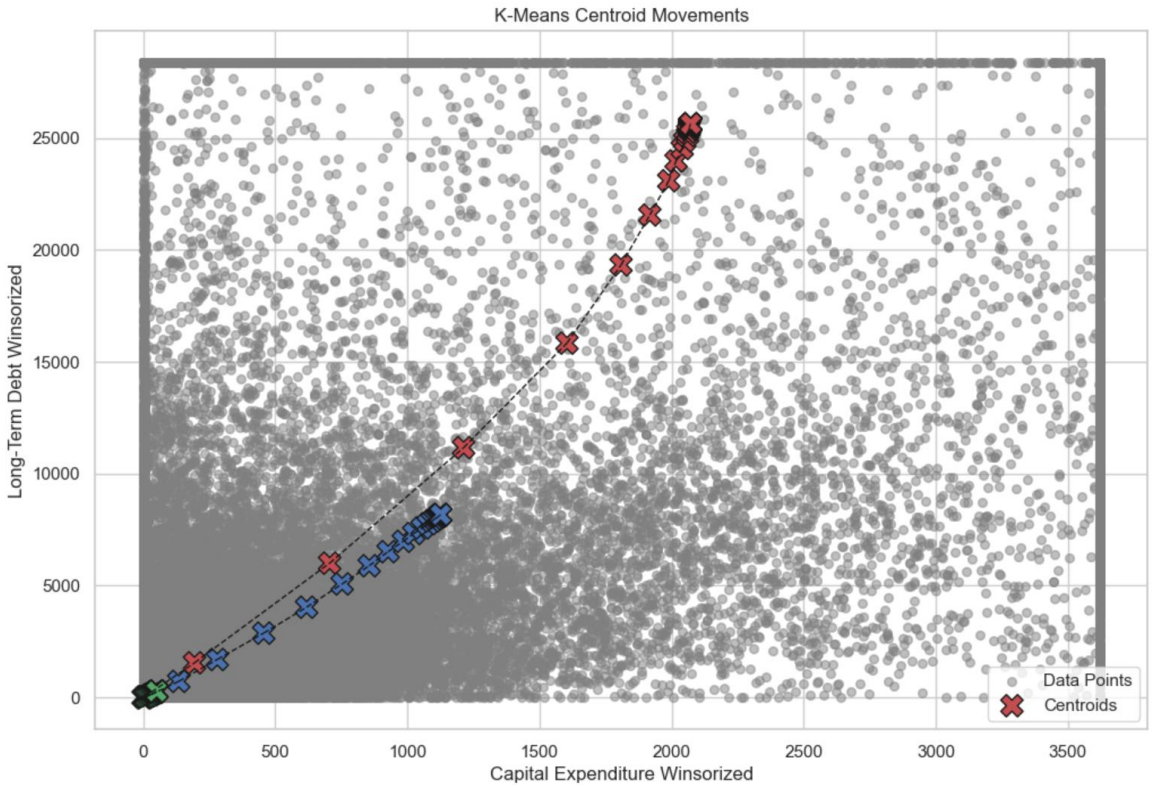
Considering that we will be plotting on a two-dimensional plot, we first select our two features (fields) for visualization, “capx” and “dltt”, which are Capital Expenditure and Long-Term Debt for our financial dataset which comprises 26,154 companies from fiscal year 2000-2024. The target variable will be “sic” or Standard Industry Classification Code. We also Winsorized the fields “capx” and “dltt” using the values for 1st percentile and 99th percentile to control for outliers. The resulting table shows the results of such winsorization:

	capx	capx_winsorized	dltt	dltt_winsorized
count	235996.000000	235996.000000	2.359960e+05	235996.000000
mean	188.024907	129.670015	1.961889e+03	1023.302956
std	1232.251539	485.490747	3.559719e+04	3730.544236
min	-3258.000000	0.000000	-2.300000e-02	0.000000
25%	0.871000	0.871000	9.670000e-01	0.967000
50%	4.891000	4.891000	2.574800e+01	25.748000
75%	23.471250	23.471250	2.406628e+02	240.662750
max	82999.000000	3621.310000	4.216909e+06	28359.860000

Without winsorization on the features when we run the k-means cluster analysis with k = 3 we obtain the following graph:



As we see that the data shows a big number of outliers, we have to Winsorized the features and once we run the k-means cluster analysis on the resulting data the plot results in the following:



This plot represents K-Means clustering (k=3) applied to Winsorized Capital Expenditure and Long-Term Debt, showing data points (gray dots), cluster centroids (red, blue and green crosses), and centroid movements (dashed lines).

Key Observations

Strong Positive Correlation Between Capital Expenditure and Long-Term Debt

The clusters form an upward curved pattern, indicating that firms with higher capital expenditures tend to have higher long-term debt. This trend is expected, as capital-intensive businesses often finance investments through long-term debt.

Three Clear Clusters Based on Financial Leverage

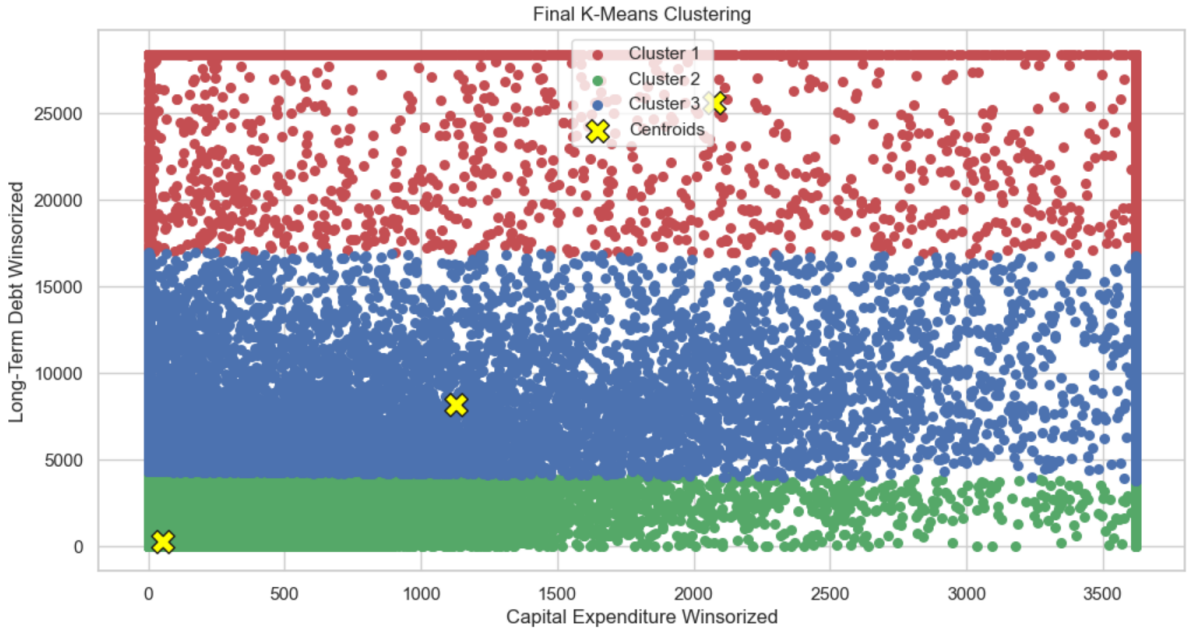
- **Green Cluster** (Lower Left Corner): Firms with low Capital Expenditure and low Long-Term Debt. Likely small firms, startups, or asset-light companies.
- **Blue Cluster** (Middle of the Curve): Firms with moderate levels of CapEx and Long-Term Debt. Likely mid-sized companies using debt financing for expansion.
- **Red Cluster** (Upper Right on the Curve): Firms with very high Long-Term Debt and Capital Expenditure. Likely capital-intensive firms (e.g., energy, manufacturing, infrastructure) that rely on heavy long-term financing.

Convergence in clusters happened after 26 iterations

Winsorization Has Controlled Extreme Values

Compared to raw data clustering (previous plot), winsorization has restricted extreme values. The distribution appears tighter, making clustering more effective.

The following plot provides the final k-means clustering with k = 3 for features Capital Expenditure Winsorized and Long-Term Debt Winsorized. The clusters are easily identifiable by green, blue and red colors.



Key Observations

Straight Horizontal Cluster Boundaries: The clusters are separated by straight horizontal lines. This suggests that Long-Term Debt dominates the clustering process, as K-Means uses Euclidean distance and prioritizes the feature with the larger range. Long-Term Debt has a much wider scale than Capital Expenditure, making it the primary factor in determining clusters. K-Means minimizes intra-cluster variance, which results in clusters that are mainly stratified by debt levels.

Key Takeaways

- K-Means effectively groups firms into three segments based on financial leverage.
- Long-Term Debt dominates clustering due to its larger scale, causing straight cluster boundaries.
- Clusters represent firms with different financial strategies: self-financed (green), moderate debt (blue), and highly leveraged (red).
- Winsorization improved clustering stability by controlling extreme values.