***Pinnacle Usage Segmentation of Mid-Market Customers***

***Ben Hefel***

***Eagle Point Software***

Since 1983, Eagle Point Software has focused on developing software solutions for AEC and manufacturing companies worldwide. Eagle Point offers award-winning software that drives productivity in technology users and helps organizations realize the full value of their Autodesk software and other business applications. With more than 500,000 global registered users, the company’s flagship solution, Pinnacle Series, has become the leader in Autodesk eLearning, Knowledge Capture & Sharing, and Productivity Improvement. Pinnacle Series is an end-to-end solution that enables AEC firms to use their complex software effectively.

***Introduction***

This project utilized customer segmentation to uncover meaningful patterns within the Pinnacle Series usage data of Mid-Market customers. The motivation for this analysis is to gain a better understanding of customers’ Pinnacle Series usage data instead of individual metrics and identify natural groupings that reveal observations across multiple variables. By applying the K-means clustering algorithm, this analysis aims to segment the dataset into a usage label that can then be used to compare customers, ultimately providing insights that support a better understanding of customer Pinnacle usage.

***Data***

The dataset for this analysis was built from scratch by writing SQL queries in Snowflake, integrating the customer, subscription, tenant, and login views from Eagle Point Software’s database (see Appendix A for base dataset query). These queries applied all necessary pre-processing, filtering, and metric calculations directly in SQL before exporting the results for analysis in Python.

This analysis focused on Mid-Market customers who have an active Autodesk content subscription and have more than zero Pinnacle users to ensure that the sample accurately represented current active and engaged customers. Several pre-processing steps were conducted within the query before the analysis began. Specifically, eight customers were excluded due to having multiple active Autodesk content library subscriptions, which would have introduced duplicate observations, and there was no single source of truth for which subscription was the ‘truth’. To identify these accounts, a separate query was written to detect duplicate transactions using the Maxio ID field from the transactions table (Appendix C), ensuring that only one active Autodesk content library subscription per customer was retained. Customers in onboarding stages were also filtered out. Finally, customers with more than 864 Pinnacle users were excluded as outliers, given their disproportionate scale compared to the rest of the sample.

Customers with active subscriptions were identified using transaction records and then joined tenant account information, which was subsequently linked to individual user IDs. For each customer with an active subscription, login activity was aggregated to produce engagement variables, which included total logins, the number of unique users who logged in, and the number of unique login days. From these statistics, multiple aggregations were calculated, which included average logins per day, user adoption rate (unique users with logins divided by total Pinnacle users), and usage per user (total logins divided by total Pinnacle users). These metrics provided a more normalized approach to representing the customers’ Pinnacle usage.

To standardize these metrics, each engagement measure was transformed into a percentile score using the PERCENT\_RANK() window function available in Snowflake. Four key usage metrics were measured: login volume (total number of logins), user breadth (number of unique users with logins), consistency (number of distinct days with logins), and adoption (proportion of active users relative to total Pinnacle users). These standardized scores were then combined into a weighted composite usage score named USAGE\_SCORE, where login volume was weighted at 35%, user breadth at 25%, consistency at 25%, and adoption at 15%. This composite score served as the primary input for K-means clustering.

After clustering, additional customer attributes - such as customer success manager, start date, end date, renewal date, and each account’s most recent health score - were joined to the results (see Appendix B for customer attributes query). These added details helped identify and interpret the resulting clusters and provided context for understanding differences in usage patterns across customer groups.

***Methods***

*Software and Environment*

Customer segmentation was utilized in VS Code using the Jupyter extension, which allowed Python to be run in a markdown-style notebook. The following libraries were used throughout the analysis: pandas, polars, scikit-learn, matplotlib, and datetime. Data preparation was done in Snowflake, as outlined in the *Data* section, then the resulting csv from the snowflake query was read into Python using pandas.

*Preprocessing*

Before clustering, the selected features were standardized using StandardScaler, from scikit-learn’s preprocessing package, to transform them to zero mean and variance, i.e., a normal distribution. This is why the filter for less than 864 pinnacle users was implemented, since StandardScaler can be sensitive to outliers. This was essential to the analysis, since K-means clustering is sensitive to the scale of input features, as features with larger numerical ranges would heavily influence the centroid distance calculations.

*Determination of Optimal Clusters*

To determine the optimal number of clusters (k), silhouette scores were calculated for k values ranging from 2 to 10. For each k, a K-means model was fitted to the standardized feature matrix, and the average silhouette score was calculated. The silhouette score measures the degree of separation between the clusters, with values closer to one indicating more distinct and cohesive clusters. The silhouette plot (*Figure 1*) shows the score for each k tested.

*Clustering Procedure*

The final K-means clustering model was fitted with k = 2 on the standardized feature dataset. Cluster assignments were added back to the original dataset for interpretation. To create interpretable labels, clusters were ranked according to their mean USAGE\_SCORE, which is defined in the *Data* section. Clusters were ranked by mean usage score and labeled accordingly. A scatter plot of Pinnacle Usage versus Usage Score (*Figure 2*) was created to visualize the clusters.

*Post-Clustering*

After clustering, the clusters were mapped back to the original dataset, and additional attributes from the *Customers* tables, which were described in *Data*, were joined to the segmentation table. For each customer, the most recent health score was used, along with columns from the *Customer* table, which included ‘CUSTOMER\_SUCCESS\_MANAGER’, ‘ACTIVE\_ARR’, ‘TOTAL\_SCORE’, and ‘RENEWAL\_DATE’. These columns allowed for more insight into the customer's data.

            The next step was to do some investigation into subgroups of the low usage cluster. The subgroups were as follows:

* Low usage & low health: Usage Score < 50 and Health Score < 50.
* Low usage & high health: Usage Score < 50 and Health Score > 50.
* High usage & high health: Usage Score ≥ 50 and Health Score ≥ 50.

Accounts assigned to “Tech Touch” were excluded from these groups to focus on accounts with an assigned Customer Success Manager.

*Renewal Filtering*

To focus on customers with upcoming renewals, the dataset was filtered to only include accounts with renewal dates before the end of the year. This enabled easier identification of which customers may need attention soon, since their renewal dates are approaching. Limiting the timeframe also reduced the chance of including customers whose renewals are too far away to take meaningful action on now.

***Results***

*Selecting the Number of Clusters*

Silhouette analysis compared models for k = 2 through k = 10. The maximum silhouette score was achieved at k = 2 (0.615), indicating that a two-cluster solution was most appropriate for this dataset (Figure 2).

*Cluster Makeup*

The final model included 279 mid-market customers, while 228 were classified as low usage, and 51 were classified as high usage. Average usage scores were 74% for High Usage and 43% for Low Usage. Cluster Centroids are shown in *Table 1*.

*Table 1. Cluster Centroids*

**Cluster Pinnacle Users Total Logins Unique Users with Logins**

Low Usage 131.96 464.08 38.72

High Usage 434.31 2836.59 148.73

*Cluster Visual*

A scatter plot of Pinnacle Users vs. Usage Score shows a split between clusters (Figure 2). High Usage points are concentrated at higher usage scores, while Low Usage points are lower on the same axis.

*Cross Examination*

After joining the *Customer* attributes, the average health score was 69.03% for low usage, and 80.82% for high usage. The usage score and health score were then used to create three different groups within the low usage table, as defined in *Methods.* (Table 2).

*Table 2. Cross Examination of Usage and Health Score*

**Group**  **Count**

Low Usage & Low Health 13

Low Usage & High Health 109

Higher Usage & Higher Health 98

*Renewal Focus*

Filtering renewals before the end of year returned 58 customers. Of these 58 customers:

* 4 were Low Usage & Low Health.
* 23 were Low Usage & High Health.

These are shown in Figure 3, which plots usage score vs. health score for the renewal group.

These results provide a clear picture of how Mid-Market customers differ in their Pinnacle Series usage patterns and how those patterns relate to health scores and renewal timelines. While the segmentation highlights which customers may need the most immediate attention, it also raises important questions about the factors influencing the customers’ health score. The following discussion explores these findings in more depth, their implications for customer retention, and opportunities for refining the analysis.

***Figure References***

*Figure 1. Silhouette scores by k (2-10)*

A graph with blue lines and dots

AI-generated content may be incorrect.

*Figure**2. Pinnacle Users vs. Usage Score colored by cluster*

***A diagram of orange and blue dots

AI-generated content may be incorrect.***

*Figure 3. Renewal-near customers: Health Score vs. Usage Score.*

*A graph of blue dots

AI-generated content may be incorrect.*

***Discussion***

The customer segmentation results show a clear separation between high and low Pinnacle usage among Mid-Market customers, with the high usage group displaying notably higher average health scores. This suggests a strong relationship between consistent usage and overall health scores. The existence of customers with lower usage scores and higher health scores reveals a discrepancy in this analysis. These customers are more than likely to benefit from other factors such as their Customer Success Manager relationship, when calculating their health score. Understanding these factors can help refine how usage data is used to generate an overall score for these customers.

Looking at this analysis from a business perspective, identifying low usage customers – particularly ones with upcoming renewal dates – is essential for retention efforts. Accounts in the low usage and low health group are at a higher risk of churn rate. Recognizing these customers in advance allows CSMs to take initiative, whether that be more meetings, additional training, or adoption initiatives to be implemented. This same approach still applies to customers with lower usage scores and higher health scores. These customers may not be as high of a churn risk, but it is still imperative that these customers receive the appropriate amount of attention.

This analysis was limited to Pinnacle usage data and did not incorporate other product activity or qualitative insights from CSMs. The composite Usage Score weights were created based on assumptions and could be adjusted based on more conversations with the right people.

***Conclusion***

This analysis used K-Means clustering on a composite usage score derived from different engagement metrics to segment mid-market customers into high and low pinnacle usage groups. The results showed that the low usage group had more customers with low health scores and upcoming renewals, making them a higher priority for attention. The goal of the analysis was to understand patterns in Pinnacle usage and identify accounts that might need extra attention before their renewals. Future work could include adding more pinnacle usage data, tracking usage changes over time, and building a model to predict which customers might be on track to move into the low usage group.

***Appendix***

***Appendix A – Usage Dataset Query***

The following SQL query was written from scratch to build the base dataset used in this analysis. It integrates data from Eagle Point Software’s Snowflake environment, performing all necessary filtering, joins, and calculated fields before exporting for analysis.

<https://github.com/benhefel8/Cluster-Analysis/blob/main/Usage%20Query.txt>

***Appendix B – Customer Attributes Query***

This second query was written to pull customer-level metadata and the most recent health score for each account, to be joined with the cluster output for interpretation.

<https://github.com/benhefel8/Cluster-Analysis/blob/main/Customers%20Query.txt>

***Appendix C – Duplicate Maxio ID Detection Query***

The following SQL was written to detect customers with more than one active Autodesk Content Library transaction at a time, based on their Maxio ID.

<https://github.com/benhefel8/Cluster-Analysis/blob/main/Duplicates%20Query.txt>