Maine Trust for Local News Subscriber Analysis

Kaylee Faherty, Cai Peng, and Sunni Raleigh

DS5110: Essentials of Data Science

Abstract

This report presents an analysis of subscriber data from the Maine Trust for Local News (METLN), conducted to support strategic efforts in subscriber retention and acquisition. The project aimed to uncover patterns in subscription behavior and identify actionable insights that could inform future engagement strategies. Using a combination of statistical summaries and clustering techniques, the characteristics of current subscribers were examined, with a particular focus on subscription length and historical engagement. The findings offer a clearer picture of the subscriber base and suggest targeted approaches to improve retention and expand readership.

Contents

1 Introduction

The Maine Trust for Local News serves a diverse readership across Maine, and understanding its subscriber base is critical to sustaining and growing its impact. This project was designed to analyze existing subscriber data from February to October 2024 to uncover meaningful trends that inform actionable recommendations. The analysis focused on two key areas: a statistical overview of the subscriber population and a clustering-based segmentation of subscribers by their subscription history. By examining variables such as subscription length, engagement patterns, subscriber locations, and demographic indicators, we sought to identify distinct subscriber groups and understand what differentiates long-term subscribers from those at risk of cancellation. These insights are intended to guide future outreach, product development, and marketing strategies. The report concludes with recommendations on the next steps, including potential areas for deeper analysis and opportunities to

enhance subscriber experience through data-driven decision-making.

2 Background

The Maine Trust for Local News, like many publishers nationwide, is experiencing a significant shift in their subscriber base from traditional print formats to digital subscriptions. While digital readership has expanded rapidly, this transition has introduced new challenges for the organization's revenue model. This shift is also reshaping the customer base, requiring new strategies to sustain subscriber retention and engagement. These dynamics provide important context for the analysis conducted in this project.

3 Methods and Analysis

3.1 Data Overview

The dataset provided for this project includes subscriber records from the Maine Trust for Local News over a ten-month period, spanning February through October 2024. In total, the data covers more than 60,000 unique subscribers across 537 cities and towns in Maine, offering a comprehensive view of the organization's readership base. Each of the ten monthly files contains columns with information such as account ID and location, along with subscription details including publication name, format (digital or print), billing method, and rate code. Additional fields capture the month and year for which a subscriber originally started their subscription, as well as the most recent date when they restarted after pausing.

While the dataset provides a strong foundation for understanding subscriber behavior, its raw form contains inconsistencies, extraneous information, and formatting issues that must be addressed before analysis. The next section outlines the steps taken to standardize the data files and compile them into a single dataset suitable for analysis. This process is essential to ensure that the data is reliable, reproducible, and ready for deeper statistical exploration.

3.2 Cleaning and Formatting

To prepare the data for deeper analysis, the information from each of the ten monthly files was combined into a single master dataset covering the entire ten-month period. Before this step, each document was re-saved from the Strict Open XML spreadsheet format (`.xlsx`) into `.csv` to ensure compatibility with the Pandas `read_csv()` function in Python.

Several additional columns were generated during the cleaning process. A month and year column was created to indicate the reporting period for each entry. Tenure variables were derived, including months since original start date and months since last start date, based on the provided date fields. A binary indicator column was added to distinguish between print and digital subscriptions. The original data files included a `publication` column with coded values, so a new column was created to display the full publication names. Finally, a churn flag column was introduced to indicate whether a subscriber remained active at the dataset's final reporting period.

Beyond compiling the data and generating new variables, additional cleaning steps were applied. Out-of-state subscribers were excluded to keep the focus on Maine readership. All column names were converted to lowercase, with underscores replacing spaces to ensure consistency. To simplify the dataset, unused columns including `Dist ID`, `Route ID`, and `Legacy Acct ID` were excluded from the analysis. Zip codes were standardized by padding four-digit entries with a leading zero, correcting cases where Excel had dropped the initial digit. Any rows missing critical identifiers, such as `subscriber_id` or `publication` codes, were removed to maintain data integrity.

4 General Results

This section summarizes key findings from the cleaned subscriber dataset. Churn rates were calculated across several dimensions, including publication, format, billing method, and geographic distribution. Confidence intervals and statistical tests were used to assess the significance of observed differences.

4.1 Overall Churn Rate

For the churn rate calculations, churn was defined as the proportion of subscribers no longer active at the final reporting period. This does not include subscribers who may have canceled and restarted their subscriptions during the 10-month period. This definition of churn rate avoids double counting people who canceled and restarted. The overall churn rate was 2% (95% CI: 1.9%-2.1%), with 1,199 subscribers no longer active by the final month.

4.2 Churn by Publication

Churn rates varied across publication. The Times Record had the highest churn at 3.4% followed by the Morning Sentinel (2.5%) and Sun Journal (2.2%). The Portland Press Herald/Maine Sunday Telegram, which accounted for the largest share of subscribers, showed a lower churn rate of 1.8% while the Kennebec Journal had the lowest at 1.3%. These differences were statistically significant ($p < 0.001$), though effect sizes were modest (Cramer's V = 0.024).

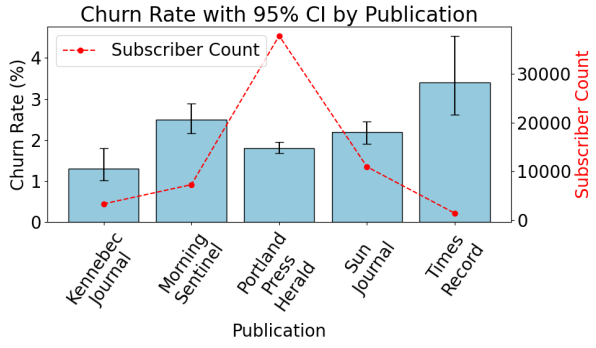


Figure 1: Churn Rate with 95% Confidence Interval by Publication

4.3 Churn by Format

Digital subscribers exhibited higher churn (2.2%) compared to print subscribers (1.7%). Confidence intervals confirmed this difference as statistically significant ($p < 0.001$). This suggests that retention challenges are more pronounced among digital readers.

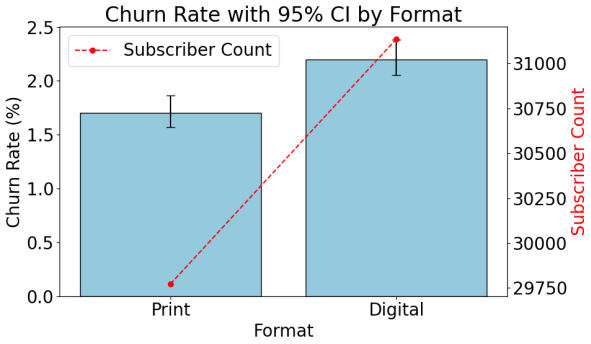


Figure 2: Churn Rate with 95% Confidence Interval by Format

4.4 Churn by Billing Method

Billing method was strongly associated with churn. Office Pay subscribers had the highest churn (2.9%), while Auto Pay - Credit Card subscribers had significantly lower churn (1.5%). Other billing methods such as Auto Pay - Bank Draft, Unpaid Comp, and Paid Comp showed even lower churn, although the number of subscribers with these billing methods is significantly less than those with Office Pay or Auto Pay - Credit Card. The differences in churn by billing method were highly significant ($p < 0.001$) and showed a moderate effect size (Cramer's V = 0.052).

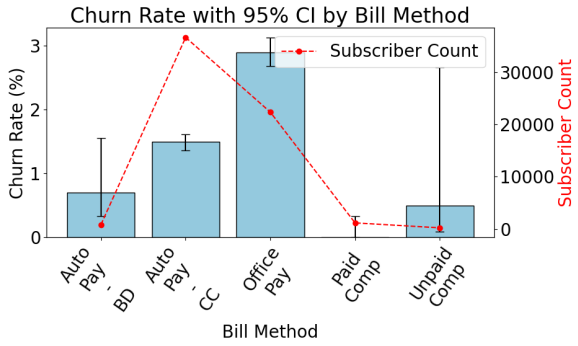


Figure 3: Churn Rate with 95% Confidence Interval by Bill Method

4.5 Geographic Distribution

Although most of the cities had low churn, several small towns with at least 10 subscribers showed elevated rates. Troy and Bucksport had churn rates of 20%, while Stoneham followed at 18.2%. These elevated rates should be interpreted cautiously given the small sample sizes, which also lead to wide confidence intervals. However, overall geographic variation was statistically significant ($p < 0.001$).

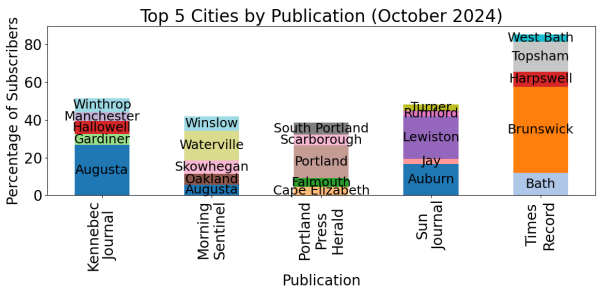


Figure 4: Top 5 Cities by Publication

4.6 Rate Code Behavior

Certain subscription rate codes were associated with extremely high churn. Trial-based codes such as `O7DAY_TRIAL10S1` (36.6%) and `SJ_O7DAY_TRIAL$10S1` (35.6%) showed the highest attrition, while standard or legacy codes had much lower churn. This pattern indicates that promotional offers attract short-term subscribers who are less likely to retain, indicating a need to balance acquisition campaigns with retention strategies. Differences across rate codes were highly significant ($p < 0.001$), with a large effect size (Cramer's V = 0.22)

4.7 Summary of Statistical Testing

Chi-square and z-tests confirmed that churn rate differences across publication, format, billing method, city, and rate code were statistically significant ($allp < 0.001$). Effect sizes ranged from small (publication, format) to large (rate codes), highlighting that while some differences are modest, others reflect substantial variation in subscriber retention. While statistically significant, these differences are small in practical terms, suggesting that format and publication are less predictive of churn compared to billing method or rate code.

5 Clustering Results

5.1 Cluster Identification

To segment subscribers and identify distinct behavioral patterns, k-means clustering was applied to the October 2024 active subscriber data. The analysis incorporated both numeric and categorical features that capture subscription behavior and preferences. Three numeric features were selected: tenure from original start date (measuring overall customer lifetime), age of current term (measuring recency of the current subscription period), and renewal gap (measuring interruptions in service). Three categorical features were included: channel preference (Digital vs. Print), bill method grouping (Auto Pay vs. Office Pay/Other), and publication name. Categorical variables were one-hot encoded prior to clustering, and all features were standardized to ensure equal weighting in the distance calculations.

The k-means algorithm was configured to identify three distinct clusters, which provided an interpretable segmentation of the subscriber base while maintaining sufficient granularity to inform targeted strategies. The resulting clusters are visualized in Figure using the first two principal components, which capture the major axes of variation in the subscriber data.

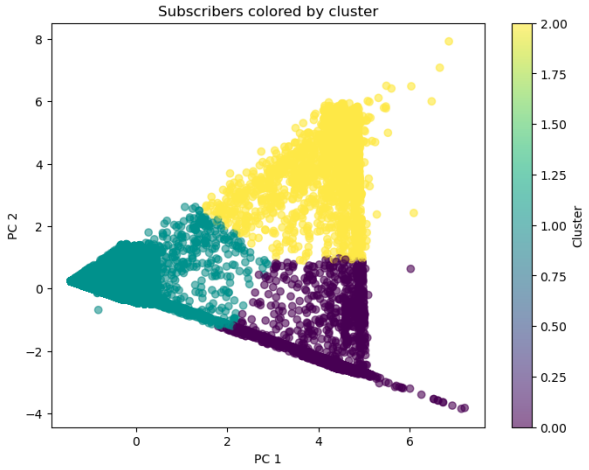


Figure 5: Subscriber segmentation visualized in principal component space. Three distinct clusters are identified using k-means clustering.

5.2 Cluster Analysis

The three identified clusters exhibit markedly different characteristics across subscription history, channel preferences, and engagement patterns.

5.3 Cluster 0: Core Long-Term Print Subscribers

Cluster 0 represents the core long-term print subscribers, comprising the largest segment of the subscriber base. As shown in Figure 6, these subscribers have maintained relationships with the Trust for approximately 11,000 days (roughly 30 years), with their current subscription terms averaging a similar duration of 10,500 days. The minimal renewal gap of approximately 400 days suggests these are highly loyal subscribers with continuous engagement.

Figure 7 reveals that approximately 85% of Cluster 0 subscribers access content through print channels, with only about 15% using digital. Figure 8 demonstrates that nearly all Cluster 0 subscribers read the Sun Journal, indicating strong geographic concentration. Payment preferences in Figure 9 show that about 80% use office pay methods, with only 20% enrolled in auto-pay credit card billing.

5.4 Cluster 1: Digitally-Engaged Moderate Tenure Subscribers

Cluster 1 represents digitally-engaged subscribers with moderate tenure. Figure 6 indicates these subscribers have been with the Trust for approximately 1,500 days (about 4 years), with current terms averaging 1,300 days and minimal renewal gaps of around 200 days, suggesting consistent engagement without significant lapses.

Figure 7 shows this cluster is distinctly digital-forward, with approximately 60% accessing content through digital channels compared to 40% through print—the highest digital penetration across all clusters. The publication distribution in Figure 8 shows greater diversity than Cluster 0, with the Portland Press Herald/Maine Sunday Telegram capturing nearly 20% of subscribers, the Morning Sentinel about 13%, and smaller representations from Kennebec Journal and Times Record. Figure 9 reveals the highest auto-pay adoption rate across all clusters, with approximately 67% using auto-pay credit card billing, suggesting stronger digital payment integration.

5.5 Cluster 2: At-Risk Reactivated Subscribers

Cluster 2 identifies a concerning at-risk subscriber segment characterized by interrupted engagement. Figure 6 shows these subscribers have total tenure of approximately 11,000 days, similar to Cluster 0, but with dramatically shorter current terms averaging only 1,800 days. Most critically, this cluster exhibits an extremely high renewal gap of approximately 9,000 days (nearly 25 years), indicating these subscribers had long periods of inactivity before recently resubscribing.

Figure 7 shows this cluster leans heavily toward print (about 80%), similar to Cluster 0, with only 20% digital. The publication distribution in Figure 8 shows dominance by the Sun Journal (approximately 70%) with signif-

icant representation from the Portland Press Herald/Maine Sunday Telegram (about 30%). Payment methods in Figure 9 show approximately 73% using office pay, with 27% on auto-pay and minimal representation of other payment types.

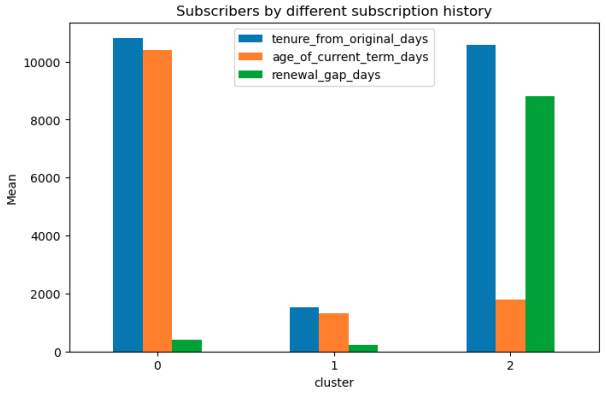


Figure 6: Subscription history metrics by cluster. Cluster 0 shows long continuous tenure, Cluster 1 shows moderate recent tenure, and Cluster 2 shows interrupted engagement with large renewal gaps.

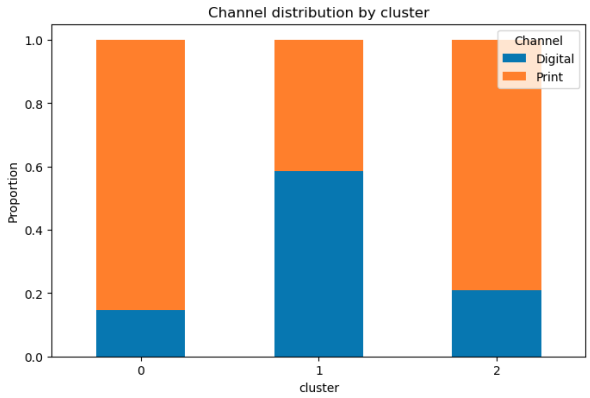


Figure 7: Channel distribution by cluster. Cluster 1 exhibits the highest digital adoption rate at approximately 60%, while Clusters 0 and 2 are predominantly print-oriented.

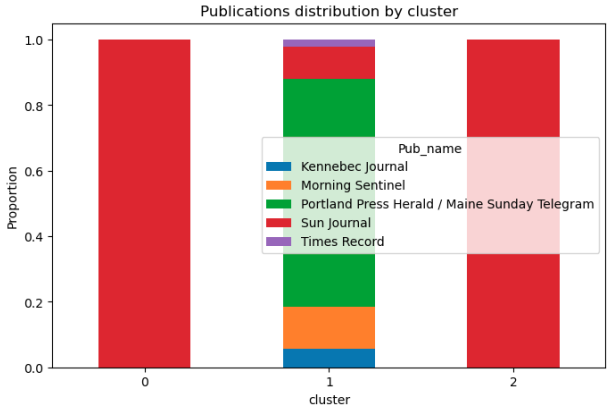


Figure 8: Publications distribution by cluster. Cluster 0 is almost exclusively Sun Journal readers, while Cluster 1 shows greater diversity across publications.

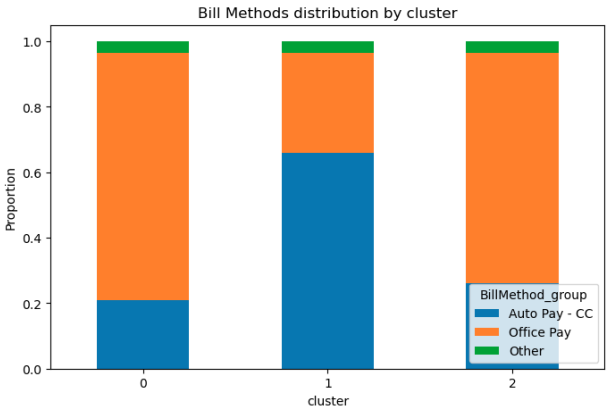


Figure 9: Bill methods distribution by cluster. Cluster 1 demonstrates the highest auto-pay adoption at 67%, while Clusters 0 and 2 predominantly use office pay methods.

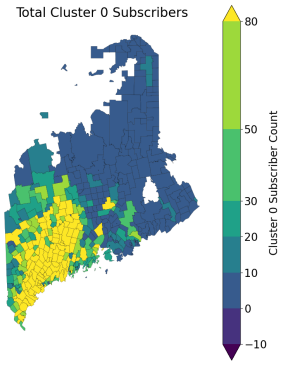
5.6 Strategic Implications

These three clusters represent distinct strategic opportunities: Cluster 0 represents the stable legacy subscriber base that requires retention focus, Cluster 1 represents the growing digital future with potential for expansion, and Cluster 2 represents a win-back population that has demonstrated renewed interest but may require additional involvement to prevent future churn.

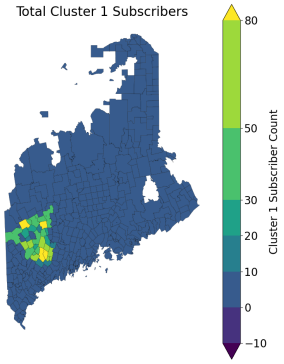
5.7 Visualizing Clusters Geographically

The following visualizations show where the subscribers in each cluster exist.

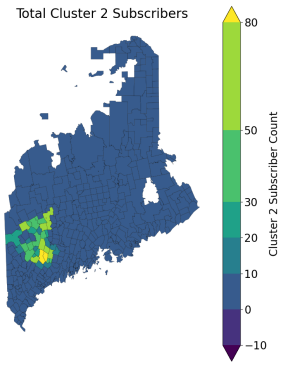
For cluster 0, the most loyal subscribers spread across Southern Maine.



For cluster 1, the newer subscribers show up in the Maine Lakes region.



Finally, for cluster 2, the win-back subscribers are located similarly in the Maine Lakes region.



6 Discussion

Overall, we see that most of METLN's subscriber-ship is in Southern Maine. When clustering subscribers, there are three distinct groups defined by their loyalty/tenure with METLN. The most prolific cluster is the first one: the most loyal subscribers. These members have been subscribed for decades and are spread throughout Southern Maine. Although they are the most geographically diverse, their readership is solely around the Sun Journal, the publication METLN has the longest relationship with. This suggests that the most loyal subscribers are the oldest, meaning METLN subscriber-ship is strongest where the relationship between publication and member is strongest/longest.

The other two clusters center around the Maine Lakes Region and show a budding readership of METLN's newest and largest publications.

These clusters can be used to group subscribers and find like members and future subscribers.

7 Limitations

Although we manipulated the date fields to create a tenure-ship-like column for our clusters, the date fields lacked nuance and required further mutations for better insights and measurements.

Additionally, subscriber demographics were limited to the region/zip code of a member. It would be nice to include more detailed demographic data like age, income/job, and gender to both strengthen our clusters and provide more robust information to our insights.

8 Future Work

Our clusters would be useful combined with revenue and marketing data. If combined with that information, we could analyze relationships between marketing campaigns, subscriber clusters, and revenue streams, creating further revenue generating actionable insights.

Additionally, these clusters could support the build of an exploratory dashboard or used as a base group to do more analysis on. If we were given more time, we would dig into the new members cluster to better understand what compels these subscribers to join.