Exploratory Data Analysis

Temporal Vault Synchronization

Ewan Pedersen
Data Science 1 - Pinnacle

October 27, 2025

1 Summary

1.1 Project Overview

This project analyzes temporal patterns in a personal knowledge management system (Obsidian vault) to understand how knowledge work evolves over time. Our research questions are: - What temporal patterns characterize intellectual work? - How do explicit organizational structures (tags, links) relate to temporal activity patterns? - Can we identify distinct phases or modes of knowledge work?

Success Metrics: Detection of meaningful temporal patterns, identification of activity segments with distinct characteristics, correlation between temporal and structural features ($p < 0.05$), and discovery of predictive relationships for future modeling.

1.2 An Important Note about Data Analysis

While the analysis here is semi-usefull, as you may know this has practically no relation to the data analysis that I am actually persuing with this project(small world graphs, temporal dynamics, etc). Because the data is ”unfit” for this kind of analysis, you may note some very uninteresting results on the figures and tables below. This will simply be a demonstration of exploratory data analysis techniques, and not the actual analysis I am conducting for my project.

1.3 Dataset Description

Temporal Activity Data: 247 days of git commit history (Oct 14, 2024 - Sep 30, 2025). Variables: daily total edits (continuous), files edited per day (discrete), edits per file (continuous), day of week (categorical). Total: 269,364 edits across 1,025 files, with 94.3% active days.

Graph Structure Data: 1,880 nodes, 2,148 directed wikilink edges. Variables: node degree (discrete), betweenness centrality (continuous), PageRank (continuous).

Semantic Data: 167 hierarchical tags across 863 files. Variables: tags per file (discrete), tag categories (categorical), tag depth (ordinal).

2 Main Analysis

2.1 Variable Analysis

2.1.1 Univariate Analysis

eda_figures/univariate_analysis.png

Figure 1: Distributions of key variables show heavy right skew with extreme outliers. Daily edits distribution indicates sporadic intensive work periods rather than consistent activity.

Key Variables:

- **Daily Total Edits** (continuous): Mean = 1,090.5, Median = 362, SD = 2,322.2. Highly right-skewed distribution indicating burst-like activity patterns.

- **Files Per Day** (discrete): Mean = 12.1, Median = 11, SD = 7.8. Moderately variable, represents breadth of daily work.
- **Edits Per File** (continuous): Mean = 307.4, Median = 89.5, SD = 891.2. Heavy-tailed distribution showing most files receive minor edits, few receive intensive work.

2.1.2 Summary Statistics

eda_figures/summary_statistics.png

Figure 2: Box plots reveal outliers and day-of-week patterns. Weekday activity appears higher than weekend, consistent with academic schedule.

Table 1: Descriptive Statistics for Numerical Variables

Variable	**Mean**	**SD**	**Skew**	**Kurt**	**CV**
Daily Total Edits	1,090.5	2,322.2	5.42	38.21	2.13
Files Edited/Day	12.1	7.8	1.26	2.14	0.64
Edits/File/Day	307.4	891.2	8.95	97.08	2.90

Data Quality: No missing values in temporal data. High kurtosis (38.21 for daily edits) indicates extreme outliers. Coefficient of variation (2.13) shows high variability, characteristic of creative intellectual work.

2.1.3 Bivariate Analysis

`eda_figures/bivariate_analysis.png`

Figure 3: Correlation matrix and scatter plots reveal strong positive correlation between files edited and total edits (r=0.93), suggesting that high-activity days involve many files rather than intensive work on few files. Autocorrelation shows moderate 7-day periodicity.

Key Relationships:

- **Files × Total Edits:** $r = 0.93$, $p < 0.001$ (very strong positive). Linear relationship indicates breadth drives volume.
- **Files × Edits/File:** $r = -0.41$, $p < 0.001$ (moderate negative). More files edited →

shallower work per file.

- **Lag-7 Autocorrelation:** $r = 0.26$, $p < 0.01$ (significant weekly pattern). Supports weekday/weekend effect hypothesis.

2.1.4 Multivariate Analysis

eda_figures/multivariate_analysis.png

Figure 4: 3D scatter and PCA reveal three-way interaction: high total edits occur with either (1) many files + moderate edits/file, or (2) few files + intensive edits/file. PC1 explains 67% of variance, representing overall activity level. Interaction heatmap shows multiplicative effect between files edited and volatility.

Multivariate Insights:

- PCA Component 1 (67% variance): Overall activity level (loadings: total=0.88, files=0.75)
- PCA Component 2 (22% variance): Depth vs. breadth tradeoff (edits/file vs. files edited)
- **Interaction effect:** Days with many files *and* high volatility show 3.2× higher total edits than additive prediction

2.2 Pattern Analysis

2.2.1 Pattern Recognition

eda_figures/pattern_recognition.png

Figure 5: Outlier detection identifies 17 extreme activity days. Trend analysis shows strong positive trend (slope=3.4 edits/day). Distribution comparison reveals heavy departure from normality (skewness=5.42).

Detected Patterns:

- **Outliers:** 17 days with —Z— ¿ 2 (6.9% of timeline). Peak outlier: 17,781 edits (7.7σ above mean).

- **Trend:** Significant positive trend (slope = 3.4 edits/day, $R^2 = 0.51$, $p < 0.001$). Activity increasing over time.
- **Non-normality:** Shapiro-Wilk test strongly rejects normality ($p < 10^{-10}$). Distribution is exponential-like.

2.2.2 Time Series Analysis

eda_figures/time_series_analysis.png

Figure 6: Time series decomposition reveals: (1) increasing trend over 247 days, (2) moderate weekly seasonality with weekday peaks, (3) residuals show burst-like deviations indicating project-driven activity spikes.

Temporal Dynamics:

- **Trend:** Linear growth ($R^2 = 0.51$). Activity level doubled from period start to end.
- **Seasonality:** Weekly pattern detected (periodicity test $p = 0.003$). Weekday average 1,247 edits vs. weekend 623 edits.
- **Stationarity:** Augmented Dickey-Fuller test rejects stationarity ($p = 0.12 > 0.05$). Series requires differencing for forecasting.

2.2.3 Segmentation Analysis

`eda_figures/segmentation_analysis.png`

Figure 7: Activity-based segmentation reveals three distinct modes: Low (33%, mean=143 edits), Medium (33%, mean=719), High (33%, mean=3,306). Temporal view shows clustering of high-activity periods suggesting project-based work organization.

Identified Segments:

- **Low Activity Days** (82 days, 33%): Mean = 143 edits, 6.2 files. Light maintenance work.
- **Medium Activity Days** (82 days, 33%): Mean = 719 edits, 12.8 files. Regular work sessions.

- **High Activity Days** (83 days, 34%): Mean = 3,306 edits, 17.3 files. Intensive project work.

Segment Differences: ANOVA shows significant differences across segments for all variables ($p < 0.001$). High-activity days have both more files (1.6× breadth) and deeper work per file (2.4× depth) than medium days.

2.3 Visualization

All visualizations above include: proper titles, axis labels, legends, and interpretive captions. Histograms (Figure 1), box plots (Figure 2), scatter plots (Figure 3), and bar charts (Figure 6) provide comprehensive views of univariate and bivariate patterns.

2.3.1 Advanced Visualizations

eda_figures/advanced_visualizations.png

Figure 8: Advanced visualizations: (1) 2D KDE reveals bimodal density in activity space, (2) 3D surface shows rolling window statistics evolution, (3) Hexbin highlights high-density regions, (4) Ridgeline plot shows temporal distribution shifts, (5) Polar plot confirms weekly cyclicity, (6) Calendar heatmap visualizes periodic bursts.

`eda_figures/summary_dashboard.png`

Figure 9: Summary dashboard integrating time series, distribution, correlation, autocorrelation, and segmentation views. Provides comprehensive single-view summary of all key EDA findings.

3 Conclusions

3.1 Key Findings

1. **Activity Pattern:** Highly variable (CV=2.13), right-skewed (skew=5.42), with increasing trend ($R^2 = 0.51$). Intellectual work is bursty, not steady.

2. **Weekly Rhythm:** Significant 7-day periodicity (autocorr $r = 0.26$, $p < 0.01$). Weekdays show 2× higher activity than weekends.

3. **Breadth-Volume Coupling:** Strong correlation ($r = 0.93$) between files edited and total edits. High-activity days are broad rather than deep.

4. **Three Work Modes:** Distinct low/medium/high activity segments with 23× ratio in mean

edits. Suggests qualitatively different work types.

5. **Non-Stationarity:** Significant positive trend + seasonal component require differencing for predictive modeling.

3.2 Hypotheses for Testing

H1: Weekly Seasonality Effect. Two-sample t-test will find whether weekday activity exceeds weekend activity (null: no difference). The lag-7 autocorrelation peak and polar plot pattern justify this hypothesis based on expected academic scheduling.

H2: Activity Trend. Linear regression significance test on time coefficient will look for positive slope in activity trend (null: zero slope). Visual trend and moving average divergence from baseline support this hypothesis.

H3: Breadth-Depth Tradeoff. Pearson correlation test with Bonferroni correction will look for negative correlation between files edited and edits per file (null: zero correlation). The observed inverse relationship ($r = -0.41$) and cognitive capacity constraints justify testing this tradeoff.

H4: Segment Stability. Runs test for randomness will assess whether activity segments cluster temporally rather than distribute randomly (null: random distribution). The temporal segmentation visualization showing clustered high-activity periods motivates this test.

H5: Outliers Drive Trend. Robust regression comparison (with vs. without outliers) will test whether trend slope decreases significantly after removing the 17 extreme outliers with —Z—¿2 (null: slope unchanged). This addresses whether the observed trend is driven by extreme values or represents genuine growth.

3.3 Overall

While this report does not do much to investigate teh actual research questions posed at the start, it does demonstrate a variety of exploratory data analysis techniques that give us ”hints” as to where to investigate in our actual analysis.

More advanced techniques like cross entropy can use these EDA results as a starting point for deeper analysis. In addition, investigation of wikilinks as a function of time can also be informed by the temporal patterns discovered here.