# **Multivariate Outlier Detection in Digital Bank Activity Data: A Methodological Framework**

## **Executive Summary**

Outlier detection within digital bank activity data is a critical undertaking, extending beyond mere statistical anomaly identification to encompass the detection of fraudulent activities, system malfunctions, and atypical user behaviors. A robust methodological framework for this domain necessitates a sophisticated approach that integrates statistical principles with machine learning techniques.

Key findings indicate that a purely univariate analysis is often insufficient, as many significant anomalies manifest through unusual combinations of otherwise normal individual variable values. Therefore, prioritizing multivariate outlier detection methods, particularly those resilient to uncorrelated features, is essential. Furthermore, the effective integration of categorical data into these models demands careful consideration of encoding strategies to balance information preservation, model interpretability, and computational efficiency. Finally, leveraging statistical tests like ANOVA and Kruskal-Wallis is fundamental to understanding how categorical variables influence the distribution and variability of quantitative data, thereby establishing context-specific baselines for anomaly identification.

The primary recommendations involve implementing a multi-stage process: an initial univariate screening for obvious data quality issues, followed by a thorough contextual variability analysis using appropriate statistical tests to identify group-specific behaviors. Subsequently, advanced multivariate techniques, chosen with an understanding of their sensitivity to data characteristics and encoding implications, should be applied. This integrated and iterative approach, informed by domain expertise, will lead to more precise and actionable outlier detection in complex digital banking environments.

## **Introduction to Outlier Detection in Digital Banking Data**

The landscape of digital banking is characterized by a continuous stream of transactional and activity data, encompassing a wide array of user interactions. Within this vast dataset, the identification of outliers is not merely an academic exercise but a practical imperative. Anomalies or outliers represent data points that significantly deviate from the expected behavior of a dataset.1 In the context of digital banking, these deviations can signal critical events such as fraudulent transactions, underlying system inefficiencies, or unusual customer interactions that warrant immediate attention. For instance, just as outliers in geochemical exploration can indicate mineral deposits 2, anomalies in digital banking data can pinpoint system malfunctions or other irregularities that compromise system integrity, lead to financial losses, or degrade customer experience.3

The data encountered in digital bank activity is inherently mixed-type. It typically includes quantitative variables, such as the "frequency of clicks/events" and "time taken" for specific actions or processes. These are continuous numerical measures that can exhibit diverse and often complex statistical distributions. Complementing these are categorical variables, such as "market/country," "case category," and "line of business." These discrete variables serve to classify or group the data, providing crucial contextual information. The inherent challenge in outlier detection for such datasets lies in effectively analyzing these disparate data types to uncover anomalies that might manifest either within individual dimensions or, more subtly, through unusual combinations across multiple dimensions or within specific subgroups defined by the categorical variables. Addressing this mixed-data challenge is paramount for comprehensive and accurate anomaly detection.

## **Univariate vs. Multivariate Outlier Detection: A Fundamental Distinction**

Understanding the nature of outliers is crucial before embarking on detection methodologies. Outliers are typically classified based on the number of variables considered in their identification.

### **Defining Outliers**

A **univariate outlier** is an observation that exhibits an extreme value on a single variable.4 For example, in digital banking, an exceptionally high frequency of clicks by a single user, irrespective of other factors, could be flagged as a univariate outlier. This type of outlier is often straightforward to detect using simple statistical measures.

As the complexity of data increases, so does the definition of an outlier. A **bivariate outlier** refers to an unusual score combination on two variables.6 Expanding on this, a

**multivariate outlier** is an observation characterized by unusual patterns across multiple variables, or a combination of unusual scores on at least two variables.4 A data point might not be extreme on any single variable when examined in isolation, but it becomes anomalous when its relationship or combination with other variables is considered. For instance, a moderate number of clicks combined with an unusually short time taken for a complex, multi-step banking process might not individually be extreme values. However, their joint occurrence could represent a multivariate outlier, potentially indicating automated activity or a system glitch.

### **Why Multivariate Detection is Crucial for Complex Systems**

Univariate methods, while useful for initial screening, analyze data one variable at a time, often failing to capture the broader context and intricate relationships within the dataset.3 This limitation means that univariate analysis can miss critical anomalies that only become apparent when the interplay between multiple variables is examined.8 For example, a high frequency of clicks might be normal for a "power user" in a digital bank. However, if this high frequency is coupled with an extremely short "time taken" for a series of complex transactions, that specific

*combination* of values is unusual. A univariate analysis focusing solely on clicks or time taken would likely fail to flag this as an anomaly, overlooking the critical contextual information provided by the other variable. This phenomenon underscores that an outlier is not always an extreme value on a single dimension but can be a *contextual anomaly* arising from an atypical combination of otherwise non-extreme values. This observation highlights that even if individual quantitative variables like frequency and time taken are not strongly linearly correlated, their joint distribution can still reveal significant anomalies. The absence of linear correlation does not diminish the necessity for multivariate analysis; instead, it emphasizes the importance of employing methods that do not solely rely on linear relationships.

Table 1 provides a comparative overview of univariate and multivariate outlier detection approaches.

**Table 1: Comparison of Univariate vs. Multivariate Outlier Detection**

| Feature | Univariate Outlier Detection | Multivariate Outlier Detection |

| :--- | :--- |:--- | | Definition | An extreme value on a single variable. | An observation with unusual patterns across multiple variables, or a combination of unusual scores on at least two variables. |

| Scope | Examines each variable in isolation. | Considers the interplay and relationships between multiple variables simultaneously. |

| Types of Outliers Detected | Extreme individual values (e.g., a single very high click count). | Unusual combinations of values that may not be extreme individually (e.g., high clicks with unusually short time taken). |

| Examples | Z-score > 3.29, values outside 1.5 IQR in a box plot. | Mahalanobis Distance, Isolation Forest, Local Outlier Factor (LOF). |

| Common Methods | Z-score, Interquartile Range (IQR) via Box Plots, Histogram-based methods. | Mahalanobis Distance, Isolation Forest, Autoencoders, Clustering Algorithms (DBSCAN, K-means), LOF, k-Nearest Neighbors (KNN), Principal Component Analysis (PCA) based methods. |

| Limitations | Misses anomalies that arise from the interaction of multiple variables; lacks contextual understanding. | Can be more complex computationally; requires careful handling of mixed data types and assumptions about data distribution. |

## **Strategies for Quantitative Data Outlier Analysis: Joint vs. Separate**

The choice between performing frequency-based and time-taken-based outlier analysis together (jointly) or separately depends fundamentally on the nature of the anomalies sought and the underlying data structure.

### **When to Perform Joint Analysis (Multivariate)**

It is generally advisable to favor joint, or multivariate, analysis when dealing with multiple quantitative variables, even if initial assessments suggest they are not strongly correlated.3 The rationale is that multivariate outliers are defined by their inconsistency with the overall correlational structure of the dataset.7 This includes complex, non-linear relationships, or simply unusual combinations of values that, individually, might fall within expected ranges. For instance, in digital banking, a user's click frequency and time taken for a task might not exhibit a strong linear correlation across the entire user base. However, a specific combination, such as a very high frequency of clicks completed in an exceptionally short time, could be highly anomalous for a human user, indicating bot activity or system manipulation. This type of anomaly would be missed by separate univariate analyses.

Several methods are well-suited for joint multivariate analysis:

* **Mahalanobis Distance (D²):** This is a classical multivariate technique that quantifies the distance of a data point from the centroid (mean) of the other data points, taking into account the covariance structure of the variables.2 It effectively accounts for the relationships between variables. However, the standard Mahalanobis Distance assumes multivariate normality 2 and is sensitive to the presence of outliers in its estimation of the mean and covariance matrix.2 To mitigate this sensitivity,  
  **Robust Mahalanobis Distance (RMD)** employs robust estimators for location and covariance, such as the Minimum Covariance Determinant (MCD).2 MCD identifies a subset of the data that minimizes the determinant of the covariance matrix, providing a more resilient estimation of the data's central tendency and spread.4 RMD values are then typically compared against a critical value from the chi-squared distribution to flag potential outliers.2
* **Principal Component Analysis (PCA):** PCA is a dimensionality reduction technique that transforms a set of possibly correlated variables into a new set of uncorrelated variables called principal components.9 These components capture the most significant variations in the data. Outliers can then be identified by examining the scores of data points on these principal components 9 or by applying distance-based methods like Mahalanobis Distance within the PCA-transformed space.3 This approach is particularly robust when the underlying data structure is complex or when dealing with high-dimensional data.
* **Isolation Forest:** This ensemble method operates by building random decision trees that recursively partition the data space. Anomalies, being "few and different," are typically isolated with fewer splits than normal data points.1 A key advantage of Isolation Forest is that it does not rely on density estimations or distance calculations that are sensitive to assumptions about data distribution or correlation structure, making it highly effective and robust across varying data distributions and feature correlations.
* **Local Outlier Factor (LOF):** A density-based method, LOF measures the local density of a data point relative to its neighbors.1 Points with significantly lower density than their surrounding neighbors are flagged as anomalies. This method is less sensitive to global correlation assumptions and focuses on local deviations.
* **Hotelling's T² Test:** This is a multivariate generalization of the univariate t-test, designed to detect outliers in normally distributed multivariate data.1

### **When Separate Univariate Analysis Might be Sufficient or Complementary**

While multivariate analysis is generally superior for detecting complex anomalies, univariate analysis retains significant value for initial data exploration and identifying obvious extreme values on single dimensions.5 Methods such as Z-score 5, the standard deviation method 9, and box plots 9 are simple and effective for this purpose.

A multi-stage approach, starting with univariate checks for data quality, followed by multivariate analysis for contextual anomalies, is often optimal. This is because obvious errors, such as impossible time-taken values (e.g., negative or zero) or extremely high click frequencies that defy logical human interaction, are best identified and addressed early. These types of data entry errors or extreme values can severely distort the estimations required for multivariate measures like Mahalanobis Distance.2 Thus, an initial univariate screening can serve as a crucial data quality assurance step, ensuring that subsequent, more sophisticated multivariate analyses are not compromised by fundamental data inaccuracies.

### **Addressing the Challenge of Uncorrelated Quantitative Variables**

The user's specific concern about uncorrelated quantitative values (frequency of clicks/events, time taken) is important. Even if these variables are not linearly correlated, a multivariate outlier can still exist if their *combination* is unusual.8 For example, a user with an average number of clicks but an extremely low time taken for a complex task might be an outlier, even if clicks and time taken are generally uncorrelated in the broader population. The presence of outliers can also significantly influence correlation coefficients, potentially inflating or decreasing them.12 This means that an observed low correlation might itself be a symptom of existing outliers.

Techniques particularly suitable for handling quantitative features, regardless of their correlation, include:

* **Isolation Forest:** As previously discussed, this method's mechanism of isolating anomalies through random partitioning makes it highly effective irrespective of feature correlation or data distribution assumptions.1
* **PCA-based methods:** By transforming correlated variables into a new set of uncorrelated principal components, PCA provides a decorrelated space where outlier detection can be performed effectively.3
* **Local Outlier Factor (LOF) and K-Nearest Neighbors (KNN):** These distance-based methods assess the local density or distance to neighbors.1 They identify points that are unusually far from their closest neighbors, making them less reliant on global correlation assumptions.

Table 2 summarizes key multivariate outlier detection methods for quantitative data, highlighting their mechanisms, assumptions, and suitability for various data characteristics.

**Table 2: Key Multivariate Outlier Detection Methods for Quantitative Data**

| Method | Mechanism | Assumptions/Sensitivity | Suitability for Uncorrelated Features | Strengths | Weaknesses |
| --- | --- | --- | --- | --- | --- |
| **Mahalanobis Distance (Standard)** | Measures distance from centroid, accounting for covariance. | Assumes multivariate normality; highly sensitive to outliers in mean/covariance estimation. | Yes, accounts for covariance structure. | Intuitively understandable, easy to calculate. | Sensitive to outliers, assumes continuous data, assumes multivariate normality. |
| **Robust Mahalanobis Distance (MCD)** | Uses robust estimators (e.g., MCD) for location and covariance before calculating distance. | Less sensitive to outliers; still benefits from approximate multivariate normality. | Yes, provides more reliable estimates in presence of outliers. | Yields mean with maximum possible breakdown point; typically removes fewer observations than MVE/MCD. | May remove up to 50% of sample (MCD/MVE); generally does not have as high a breakdown point as MVE/MCD (MGV). |
| **Principal Component Analysis (PCA)-based** | Transforms data into uncorrelated principal components; outliers identified in this new space. | Assumes linearity for component extraction; effectiveness depends on variance captured by components. | Yes, by design it decorrelates features. | Reduces dimensionality; can reveal outliers in lower-dimensional space. | Interpretation of components can be challenging; sensitive to scaling. |
| **Isolation Forest** | Builds random decision trees to isolate anomalies with fewer splits. | Few explicit distributional assumptions. | Yes, effective regardless of correlation structure. | Effective for high-dimensional data; computationally efficient; does not rely on distance or density. | Can be sensitive to parameter tuning (e.g., contamination); may not perform well on clustered anomalies. |
| **Local Outlier Factor (LOF)** | Measures local density of a point relative to its neighbors. | Assumes that normal data points are denser than outliers. | Yes, focuses on local relationships rather than global correlations. | Effective for detecting outliers in varying density regions; does not assume global distribution. | Computationally intensive for large datasets; sensitive to parameter 'k' (number of neighbors). |
| **Hotelling's T² Test** | Multivariate extension of t-test, compares multivariate mean to a known mean or between groups. | Assumes multivariate normality. | Yes, accounts for multivariate mean and covariance. | Direct hypothesis testing for multivariate mean differences. | Sensitive to violations of normality and outliers. |

## **Integrating Categorical Data into Outlier Detection Models**

The presence of categorical data alongside quantitative data in digital bank activity records introduces specific challenges for outlier detection, particularly when employing machine learning models.

### **The Necessity of Encoding Categorical Attributes for Machine Learning Models**

Machine learning models, including powerful anomaly detection algorithms like Isolation Forest, are fundamentally designed to operate on numerical inputs.14 They cannot directly interpret text or discrete categorical values (e.g., "market/country," "case category"). Therefore, converting these categorical attributes into a numerical representation is a mandatory preprocessing step before feeding them into such models.14

### **Detailed Discussion of Encoding Strategies for Isolation Forest**

The choice of encoding strategy significantly impacts model performance and interpretation.

* **Label Encoding:**
  + **Advantages:** This method is simple and assigns a unique integer to each category (e.g., "USA" = 1, "UK" = 2, "India" = 3).14 It is computationally efficient and keeps the dimensionality of the dataset low.
  + **Disadvantages:** A significant drawback is that Label Encoding imposes an artificial ordinal relationship between categories. For nominal data, such as 'market/country' or 'case category', this implies that 'country A' is "greater than" or "less than" 'country B', which is semantically incorrect and can mislead the model. This spurious ordering can negatively impact the model's ability to learn meaningful patterns and detect true anomalies.14
  + **When Advisable:** Label Encoding is only advisable for truly ordinal categorical data where an inherent, meaningful order exists (e.g., 'low', 'medium', 'high' risk categories).14
* **One-Hot Encoding:**
  + **Advantages:** This strategy addresses the artificial ordinality problem by creating new binary (0 or 1) features for each category. For example, 'market/country' with three categories would become three new binary columns (e.g., market\_USA, market\_UK, market\_India), where only one is '1' for a given observation.14 This ensures that each category is treated independently, without implying any order.
  + **Disadvantages:** A major concern with One-Hot Encoding is the "curse of dimensionality," especially when a categorical variable has many unique values (high cardinality). This can lead to the creation of a very wide dataset, with a large number of sparse binary features.14 For anomaly detection, this increased dimensionality can make the data "loosely bound" and clusters "far apart" 14, potentially hindering the effectiveness of density-based or distance-based methods. For Isolation Forest specifically, a large number of one-hot encoded features might increase the likelihood of these features being randomly picked for splits, potentially diluting the impact and signal from the more informative numerical features.16
  + **When Advisable:** One-Hot Encoding is generally advisable for nominal categorical data with low cardinality (a small number of unique categories).15
* **Advanced Encoding Methods (e.g., Target Encoding, Embedding Layers):**
  + For large datasets or categorical features with high cardinality, the issues associated with One-Hot Encoding become more pronounced.14 Advanced encoding methods aim to find a balance between preserving information and managing dimensionality.
  + **Target Encoding (Mean Encoding):** This technique replaces each category with the mean of the target variable (e.g., anomaly score or a binary anomaly flag) for that category. It can be highly effective at capturing information but is prone to overfitting without proper regularization techniques like cross-validation or smoothing.
  + **Embedding Layers:** Commonly used in deep learning, embedding layers learn a dense, low-dimensional vector representation for each category. These learned embeddings can capture semantic relationships between categories, offering a more nuanced numerical representation than simple integer or binary encoding.14
  + **GEL Encoding:** This method has been noted to perform better than One-Hot Encoding for larger datasets in certain contexts.14

The selection of an encoding strategy involves a critical trade-off between introducing artificial order or bias (as with Label Encoding) and managing increased dimensionality (as with One-Hot Encoding). The dimensionality increase can make it more challenging for models like Isolation Forest to identify meaningful splits, or it can dilute the signal from the numerical features.14 Advanced methods strive to navigate this balance by preserving relevant information while maintaining a manageable feature space. Therefore, the most effective encoding strategy is not universal; rather, it depends on the specific characteristics of the data (e.g., cardinality, ordinality) and the chosen anomaly detection model.14 For digital bank activity data, careful experimentation with different encoding techniques is vital to optimize outlier detection performance.

### **Approaches for Mixed-Type Data Outlier Detection Beyond Simple Encoding**

Beyond simple encoding, some advanced methodologies are specifically designed to handle mixed-attribute data, aiming to avoid the potential information loss or noise that can arise from converting all variables to a single type.17 The dilemma of information loss is critical: while simple encoding makes data compatible with an algorithm, it might obscure the true nature of an outlier that manifests through the

*interaction* of categorical and numerical features. For example, an unusually short "time taken" for a transaction might only be anomalous for a specific "case category" in a particular "market." Dedicated mixed-type methods attempt to preserve or model these complex interactions more explicitly.

These methods can be broadly classified into several categories:

* **Categorized Methods:** These approaches transform numerical data into categorical data, and then apply outlier detection techniques designed for purely categorical data.19 Examples include discretization, where numerical ranges are converted into categories, or interval-based methods that segment continuous data into bins. The analysis then focuses on association relationships or frequencies within these newly formed categorical attributes.
* **Enumerated Methods:** Conversely, these methods transform categorical data into numerical data, followed by the application of numerical outlier detection techniques.19 Examples include using Naive Bayes classifiers or Link functions to map categorical attributes to latent numerical variables, which are then analyzed using numerical distance or statistical methods.
* **Combined Methods:** This common and flexible approach involves performing outlier detection separately for categorical and numerical data components, and then combining their individual outlier scores through a specific function to derive a final overall outlier score.19 This allows for the application of specialized techniques for each data type (e.g., frequency-based for categorical, distance-based for numerical) before integration.
* **Mixed Methods:** These advanced techniques process categorical and numerical data simultaneously, explicitly considering their interactions to determine the outlier value.19 Pattern-based Outlier Detection (POD) is an example, where outliers are identified based on their deviation from learned data patterns that incorporate both numerical and categorical attributes.

For digital bank activity data, while Isolation Forest with appropriate encoding is a strong starting point, exploring methods specifically designed for mixed data types could yield more accurate and interpretable results. This is particularly true for uncovering complex fraud patterns or unusual operational events where the anomaly's significance lies in the interplay between quantitative and categorical attributes.

Table 3 provides a summary of common categorical data encoding strategies and their implications for anomaly detection.

**Table 3: Categorical Data Encoding Strategies for Anomaly Detection**

| Encoding Method | Mechanism | Pros | Cons | When to Use | Impact on Isolation Forest |
| --- | --- | --- | --- | --- | --- |
| **Label Encoding** | Assigns a unique integer to each category. | Simple, low dimensionality. | Imposes artificial ordinal relationship; can mislead models for nominal data. | Truly ordinal categorical data (e.g., 'low', 'medium', 'high'). | Can introduce spurious relationships, potentially distorting split decisions if categories are nominal. |
| **One-Hot Encoding** | Creates binary features (0/1) for each category. | Avoids artificial ordinality; treats categories independently. | High dimensionality for high-cardinality features ('curse of dimensionality'); sparse data. | Nominal categorical data with low cardinality. | Increased dimensionality can dilute signal from numerical features; more features to split on, potentially requiring deeper trees. |
| **Target Encoding (Mean Encoding)** | Replaces category with mean of target variable for that category. | Captures target variable relationship; reduces dimensionality compared to One-Hot. | Prone to overfitting without regularization; requires a target variable (supervised/semi-supervised context). | High-cardinality nominal data, when a target variable is available. | Can provide a strong signal for anomaly detection if the target variable is related to outlierness; requires careful validation. |
| **Embedding Layers** | Learns dense, low-dimensional vector representations for categories. | Captures semantic relationships; effective for high cardinality; reduces dimensionality. | Requires neural network architecture; more complex to implement and interpret. | High-cardinality nominal data, especially in deep learning contexts. | Can provide rich, compact representations of categorical data, potentially improving model performance by capturing complex relationships. |

## **Assessing Variability Dependence with Categorical Variables: ANOVA and Kruskal-Wallis**

In the context of outlier detection, understanding how the variability of quantitative data is influenced by categorical variables is paramount. Statistical tests like One-Way Analysis of Variance (ANOVA) and the Kruskal-Wallis H-test serve this purpose, helping to identify if the mean or distribution of quantitative variables differs significantly across distinct categorical groups.

### **Purpose in Outlier Detection Context**

These tests are instrumental in determining if the mean (ANOVA) or median/distribution (Kruskal-Wallis) of a quantitative variable (e.g., "frequency of clicks," "time taken") varies significantly across groups defined by a categorical variable (e.g., "market/country," "case category," "line of business"). If a categorical variable significantly influences the central tendency or variability of a quantitative variable, it implies that the "normal" behavior—and consequently, what constitutes an outlier—for that quantitative variable is *different* across the various categories. For example, the "time taken" for a specific type of transaction might be inherently longer in "Market A" due to stringent regulatory checks compared to "Market B." In such a scenario, an outlier in "Market A" would be a deviation from its specific norm, not from a global average. These tests provide the statistical foundation for establishing such group-specific baselines. This means that outlier detection should ideally be performed *within* these identified groups, or at the very least, group membership should be incorporated as a contextual factor within multivariate outlier models.

### **One-Way ANOVA**

**Purpose:** One-Way ANOVA is a parametric statistical test used to determine if there are any statistically significant differences between the means of two or more independent (unrelated) groups on a continuous dependent variable.21 While it can be used for two groups, an independent samples t-test is often more appropriate for that specific case; ANOVA is typically employed when comparing three or more groups.22

**Assumptions:** For the results of a One-Way ANOVA to be valid, several assumptions must be met:

* **Dependent Variable:** The variable of interest (e.g., "frequency of clicks," "time taken") must be continuous, measured on an interval or ratio scale.21
* **Independent Variable:** The grouping variable (e.g., "market/country," "case category," "line of business") must consist of two or more categorical, independent groups.21
* **Independence of Observations:** Observations within and between groups must be independent, meaning there should be no relationship between them (e.g., different participants in each group).21
* **No Significant Outliers:** The presence of significant outliers can negatively impact the validity of ANOVA results.21 This highlights an iterative process where initial outlier detection or robust methods might precede ANOVA, or ANOVA results might prompt further investigation into group-specific outliers.
* **Approximate Normality:** The dependent variable should be approximately normally distributed for each group.21 ANOVA is considered "robust" to minor violations of this assumption.
* **Homogeneity of Variances:** The variances of the dependent variable should be approximately equal across all groups. This assumption can be formally tested using Levene's test.21 Unequal group sizes make the homogeneity assumption particularly critical.21

**Application in Assessing Variability Dependence:** By applying One-Way ANOVA to digital banking data, researchers can test whether the mean "frequency of clicks" or "time taken" differs significantly across distinct "market/country" or "case category" groups. A significant finding implies that these categorical variables are indeed drivers of differences in central tendency, indicating that different "normal" ranges exist for each group.

**Significance and Interpretation:** A statistically significant p-value (typically less than 0.05) from an ANOVA test indicates that there is a significant difference between at least two of the group means.21 However, ANOVA is an "omnibus" test; it does not specify

*which* particular groups differ from each other. To identify specific pairwise differences, post-hoc tests (e.g., Tukey HSD, Bonferroni correction) are necessary.21

### **Kruskal-Wallis H-Test**

**Purpose:** The Kruskal-Wallis H-test, often referred to as the "one-way ANOVA on ranks," is a non-parametric statistical test.24 It serves as a robust alternative to One-Way ANOVA when the parametric assumptions, particularly normality, are not met.24 This test is used to compare three or more independent groups on a continuous or ordinal dependent variable, assessing whether there are statistically significant differences between their mean ranks.24

**Assumptions:** The Kruskal-Wallis H-test has fewer stringent assumptions than ANOVA:

* **Dependent Variable:** Must be measured at the ordinal or continuous level.24
* **Independent Variable:** Must consist of two or more categorical, independent groups.24
* **Independence of Observations:** Observations must be independent across and within groups.24
* **Similar Distribution Shapes:** While it does not assume normality, if the distributions of the dependent variable have significantly different shapes across groups, the test strictly compares mean ranks rather than medians.24
* **Sensitivity to Outliers:** The Kruskal-Wallis test is notably less sensitive to outliers compared to ANOVA.26

**Application:** This test is particularly well-suited for digital bank activity data where quantitative variables like "frequency of clicks" or "time taken" might exhibit highly skewed or non-normal distributions within specific categorical groups (e.g., many users with very low activity, few with very high). It effectively assesses whether the distributions (or medians, if distribution shapes are similar) of these quantitative variables differ across categorical groups.

**Significance and Interpretation:** A significant p-value from the Kruskal-Wallis H-test indicates that at least one group's distribution (or median, if shapes are similar) is significantly different from another.24 Similar to ANOVA, it is an omnibus test, meaning it does not pinpoint which specific groups differ. Subsequent post-hoc tests are required to identify these pairwise differences.26

### **Choosing Between ANOVA and Kruskal-Wallis**

The decision between using ANOVA and the Kruskal-Wallis H-test should be guided by the following criteria:

* **Normality:** If the dependent variable within each group is normally or approximately normally distributed, ANOVA is the appropriate choice. If normality is violated, the Kruskal-Wallis test is preferred.21
* **Homogeneity of Variances:** If ANOVA's assumption of homogeneity of variances is violated, Kruskal-Wallis might be a more suitable alternative, although robust ANOVA methods or data transformations can also be considered.
* **Outliers:** Given its lower sensitivity to outliers 26, Kruskal-Wallis can be a safer choice if outliers are present and cannot be justifiably removed or transformed.
* **Sample Size:** Kruskal-Wallis is particularly valuable for smaller datasets where it is more challenging to satisfy the normality assumptions required for parametric tests.25

These tests are indispensable for understanding the underlying structure of the data. If significant differences are identified across categorical groups, it implies that outlier detection thresholds or models might need to be tailored or applied *per group*. The primary goal of the user is outlier detection, and while ANOVA and Kruskal-Wallis do not directly detect outliers, they are critical for *defining what constitutes an outlier* within specific contexts. For example, if "time taken" for a transaction differs significantly by "case category," then an "outlier" in "fraudulent cases" might be characterized by an exceptionally short time, whereas an "outlier" in "customer service inquiries" might involve an unusually long time. These tests provide the statistical basis for establishing these context-dependent "normal" ranges and variances, which then inform the multivariate outlier detection algorithms. Without this contextual understanding, a global outlier model might incorrectly flag normal behavior in a specific group as anomalous, or conversely, miss true anomalies that are only unusual within their specific group. The results of ANOVA or Kruskal-Wallis should therefore guide the segmentation of data or the inclusion of categorical variables (e.g., as interaction terms or features) in the multivariate outlier detection model, leading to more precise and meaningful outlier identification.

Table 4 provides a comparison of One-Way ANOVA and the Kruskal-Wallis H-Test.

**Table 4: Comparison of One-Way ANOVA and Kruskal-Wallis H-Test**

| Feature | One-Way ANOVA | Kruskal-Wallis H-Test |
| --- | --- | --- |
| **Type of Test** | Parametric | Non-parametric (rank-based) |
| **Dependent Variable Type** | Continuous (interval or ratio) | Continuous or Ordinal |
| **Independent Variable Type** | Categorical (2+ independent groups) | Categorical (2+ independent groups) |
| **Hypotheses** | H0: All group means are equal. H1: At least one group mean is different. | H0: All group medians (or mean ranks) are equal. H1: At least one group median (or mean rank) is different. |
| **Key Assumptions** | 1. Continuous DV.  2. Categorical IV (2+ groups).  3. Independence of observations.  4. No significant outliers.  5. Approximate normality of DV for each group.  6. Homogeneity of variances. | 1. Ordinal or Continuous DV.  2. Categorical IV (2+ groups).  3. Independence of observations.  4. Less sensitive to outliers.  5. Similar distribution shapes across groups (for median comparison). |
| **Sensitivity to Outliers** | Highly sensitive; outliers can distort results. | Less sensitive to outliers. |
| **Post-hoc Tests** | Required to identify specific group differences (e.g., Tukey HSD, Bonferroni). | Required to identify specific group differences (e.g., Dunn's test). |
| **When to Use** | Data meets parametric assumptions (normality, homogeneity of variance); focus on mean differences. | Data violates parametric assumptions (especially normality); focus on median or distribution differences; small sample sizes. |

## **Conclusion and Practical Recommendations for Digital Bank Activity Data**

Effective outlier detection in digital bank activity data requires a sophisticated, multi-faceted approach that transcends simple univariate analysis. The complexity of financial transaction data, characterized by both quantitative and categorical attributes, necessitates a methodological framework that can uncover anomalies arising from individual extreme values, unusual combinations of values, and context-specific deviations.

An integrated strategy for outlier detection in digital bank activity data should encompass the following stages:

1. **Initial Data Quality Checks (Univariate Screening):** Before applying complex multivariate models, it is prudent to perform an initial univariate screening of quantitative variables such as "frequency of clicks" and "time taken." Simple methods like Z-score analysis or box plots can effectively identify obvious data entry errors, impossible values (e.g., negative time), or extreme values that are clearly erroneous.5 Addressing these fundamental data quality issues early is crucial, as they can significantly distort the results of subsequent multivariate analyses.
2. **Contextual Variability Analysis (ANOVA/Kruskal-Wallis):** These statistical tests are invaluable for understanding the underlying structure of the data and how categorical variables (e.g., "market/country," "case category," "line of business") influence the quantitative variables. By applying One-Way ANOVA (if assumptions of normality and homogeneity of variance are met) or the Kruskal-Wallis H-test (for non-normal data or when assumptions are violated) 21, one can determine if the mean or distribution of quantitative variables differs significantly across different categorical groups. This analysis establishes group-specific "normal" behaviors and variances, which are critical for defining what constitutes an outlier within a specific context. For instance, an "abnormal" time taken for a transaction might vary significantly depending on the "case category" or "market." The findings from these tests should guide the segmentation of data or inform the design of multivariate models by incorporating categorical variables as contextual factors or interaction terms.
3. **Multivariate Outlier Detection:** Once initial data quality is assured and contextual variability is understood, advanced multivariate methods should be employed to capture anomalies that manifest through the interplay of multiple quantitative variables, even if they appear uncorrelated. Techniques like Robust Mahalanobis Distance (utilizing estimators such as MCD to mitigate sensitivity to outliers) 2, PCA-based methods 3, or Isolation Forest 1 are highly effective. Isolation Forest is particularly advantageous due to its robustness to varying data distributions and lack of reliance on explicit correlation assumptions, making it well-suited for identifying anomalies in complex, high-dimensional digital banking data.
4. **Strategic Categorical Data Integration:** The integration of categorical data into machine learning-based outlier detection models requires careful consideration of encoding strategies. For models like Isolation Forest, converting categorical attributes to numerical representations is mandatory.14 The choice between Label Encoding (suitable only for truly ordinal data), One-Hot Encoding (effective for low-cardinality nominal data but prone to dimensionality issues with high cardinality), and more advanced methods like Target Encoding or Embedding Layers (preferable for high-cardinality features or large datasets) is crucial.14 This decision involves a trade-off between introducing artificial order, managing increased dimensionality, and preserving meaningful information. Furthermore, for mixed-type data, exploring methods specifically designed to handle both quantitative and categorical attributes simultaneously, which aim to preserve the interactions between these data types, can yield more accurate and interpretable anomaly detection results.18

Outlier detection is rarely a singular, definitive process. It is inherently iterative, involving continuous cycles of detection, investigation, and refinement of models or thresholds. Domain expertise is paramount for interpreting detected outliers, differentiating between genuine anomalies (e.g., fraud) and legitimate but unusual behaviors (e.g., a power user with unique interaction patterns).8 It is also critical to acknowledge that outliers themselves can significantly influence statistical analyses and model performance.5 Therefore, their accurate identification and appropriate handling—whether through removal, transformation, or the use of robust analytical methods—are fundamental for ensuring the validity and reliability of all subsequent data analyses and decision-making processes in digital banking.

#### Works cited

1. Multivariate Anomaly Detection - GeeksforGeeks, accessed July 7, 2025, <https://www.geeksforgeeks.org/machine-learning/multivariate-anomaly-detection/>
2. A MULTIVARIATE OUTLIER DETECTION METHOD 1 ... - CiteSeerX, accessed July 7, 2025, <http://file.statistik.tuwien.ac.at/filz/papers/minsk04.pdf>
3. Multivariate Outlier Detection: A Game Changer in Understanding Complex Systems, accessed July 7, 2025, <https://medium.com/@xai4heat/multivariate-outlier-detection-a-game-changer-in-understanding-complex-systems-deaad99e79f8>
4. Distribution of Variables by Method of Outlier Detection - PMC, accessed July 7, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3389806/>
5. Univariate and Multivariate Outliers - Statistics Solutions, accessed July 7, 2025, <https://www.statisticssolutions.com/univariate-and-multivariate-outliers/>
6. medium.com, accessed July 7, 2025, <https://medium.com/@ebimsv/machine-learning-series-day-8-preprocessing-part-3-d4feb4c7ce90#:~:text=If%20we%20detect%20outliers%20based,call%20it%20multivariate%20outlier%20detection.>
7. Outlier Detection Using Machine Learning - Charles Holbert, accessed July 7, 2025, <https://www.cfholbert.com/blog/outlier-detection-machine-learning/>
8. 3.3 Multivariate Outlier Detection - MyEducator, accessed July 7, 2025, <https://app.myeducator.com/reader/web/1421a/3/or0kd/>
9. Improving Data Quality with Outlier Detection Techniques: A CRISP-DM Approach to Data Preparation | by Donato\_TH | Donato Story | Medium, accessed July 7, 2025, <https://medium.com/donato-story/improving-data-quality-with-outlier-detection-techniques-a-crisp-dm-approach-to-data-preparation-5a04a9eef844>
10. Outliers Detection in Multivariate Time Series by Independent Component Analysis | Request PDF - ResearchGate, accessed July 7, 2025, <https://www.researchgate.net/publication/6312089_Outliers_Detection_in_Multivariate_Time_Series_by_Independent_Component_Analysis>
11. 3.2 Univariate Outlier Detection - MyEducator, accessed July 7, 2025, <https://app.myeducator.com/reader/web/1421a/3/d653e/>
12. Cautions about Correlation and Regression | STAT 800, accessed July 7, 2025, <https://online.stat.psu.edu/stat800/lesson/cautions-about-correlation-and-regression>
13. (PDF) The effect of normality and outliers on bivariate correlation coefficients in psychology: A Monte Carlo simulation - ResearchGate, accessed July 7, 2025, <https://www.researchgate.net/publication/361808138_The_effect_of_normality_and_outliers_on_bivariate_correlation_coefficients_in_psychology_A_Monte_Carlo_simulation>
14. Impact of Categorical Encodings on Anomaly Detection Methods -, accessed July 7, 2025, <https://www.analyticsvidhya.com/blog/2023/01/impact-of-categorical-encodings-on-anomaly-detection-methods/>
15. Isolation Forest : Categorical data - Stack Overflow, accessed July 7, 2025, <https://stackoverflow.com/questions/54886223/isolation-forest-categorical-data>
16. SinaDBMS/IsolationForest: Anomaly detection on mixed ... - GitHub, accessed July 7, 2025, <https://github.com/SinaDBMS/IsolationForest>
17. coms.events, accessed July 7, 2025, <https://coms.events/ntts2019/data/presentations/presentation_131.pdf>
18. MIX Model for Outlier Detection in Mixed-type Data. - ResearchGate, accessed July 7, 2025, <https://www.researchgate.net/figure/MIX-Model-for-Outlier-Detection-in-Mixed-type-Data_fig1_338947470>
19. A Survey on Mixed-Attribute Outlier Detection Methods - SciSpace, accessed July 7, 2025, <https://scispace.com/pdf/a-survey-on-mixed-attribute-outlier-detection-methods-9qn1qy0oss.pdf>
20. Outlier detection in datasets with mixed-attributes - GitHub Pages, accessed July 7, 2025, <https://vu-business-analytics.github.io/internship-office/reports/report-meltzer.pdf>
21. Section 6.2: One-Way ANOVA Assumptions, Interpretation, and Write Up, accessed July 7, 2025, <https://usq.pressbooks.pub/statisticsforresearchstudents/chapter/one-way-anova-assumptions/>
22. One-way ANOVA in SPSS Statistics - Step-by-step procedure including testing of assumptions., accessed July 7, 2025, <https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php>
23. Statistics for EES and others Comparing more than two groups: Multiple testing, ANOVA and Kruskal-Wallis, accessed July 7, 2025, <https://evol.bio.lmu.de/_statgen/StatEES/anova_handout.pdf>
24. Kruskal-Wallis-Test simply explained - Datatab, accessed July 7, 2025, <https://datatab.net/tutorial/kruskal-wallis-test>
25. The Kruskal–Wallis Test | Technology Networks, accessed July 7, 2025, <https://www.technologynetworks.com/informatics/articles/the-kruskal-wallis-test-370025>
26. Kruskal-Wallis H Test using SPSS Statistics, accessed July 7, 2025, <https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php>
27. Kruskal-Wallis H Test using Stata - Laerd Statistics, accessed July 7, 2025, <https://statistics.laerd.com/stata-tutorials/kruskal-wallis-h-test-using-stata.php>
28. Why do outliers matter? - Scribbr, accessed July 7, 2025, <https://www.scribbr.com/frequently-asked-questions/why-outliers-matter/>