MoneyVis: Open Bank Transaction Data for Visualization and Beyond

Elif E. Firat [1], Dharmateja Vytla [2], Navya Vasudeva Singh [2], Zhuoqun Jiang [2], and Robert S. Laramee[2]

[1]Department of Computer Engineering, Cukurova University, Turkey
[2]School of Computer Science, The University of Nottingham, UK

Abstract

With the rapid evolution of finance technology (FinTech) the importance of analyzing financial transactions is growing in importance. As the prevalence and number financial transactions grows, so does the necessity of visual analysis tools to study the behavior represented by these transactions. However, real bank transaction data is generally private due to security and confidentiality reasons thus preventing its use for visual analysis and research. We present MoneyData, an anonymized open bank data set spanning seven years worth of transactions for research and analysis purposes. To our knowledge, this is the first real-world, retail, bank transaction data that has been anonymized and made public for visualization and analysis by other researchers. We describe the data set, it's characteristics, the anonymization process, and present some preliminary analysis and images as a starting point for future research. The transactions are also categorized to facilitate understanding. We believe the availability of this open data will be of great benefit to the research community and facilitate further study of finance.

1. Introduction and Motivation

Despite the massive volume of electronic bank transactions throughout the world and their growing importance, the exploration and analysis of such transactions presents barriers due to privacy concerns. Bank transaction data can provide valuable insight into consumer spending patterns and financial behavior. Analysis of bank transaction data can facilitate informed decisions with respect to loans, investing, or risk management. It can also assist in tracking and detecting financial fraud, money laundering, and other criminal financial activities.

There are several finance-related data sets available online which can be used for research and analysis, however, there are often access privileges or monetary fees posing barriers to their access. Furthermore, we are unable to find an open dataset specifically on retail financial transactions. The datasets we find are usually associated with corporations, aggregate transactions, or financial statements, in other words, summaries of financial data rather than individual transactions.

Real datasets containing financial data are often hidden from the public for various reasons, including data privacy and security concerns. In order to create systems that can analyze financial data, synthetic data is often used as a substitute. Synthetic data enables the development and testing of systems for fraud detection and other financial analysis without compromising sensitive information. To the best knowledge, we provide the first authentic, publicly available, retail, financial transaction data set for analysis and research purposes. The data set has gone through an anonymization process and can be accessed via a public link provide in Section 3.2 for financial data analysis. We also present some initial analysis and images to provide preliminary understanding of the bank transaction data. The data has also been categorized to facilitate exploration and analysis. We believe that this data will be very beneficial for the researcher community and will enable further research in finance. The contributions of this paper include:

- The first open bank transaction data set from an anonymized retail customer
- Manual, semi-automatic, and automatic categorizations of the financial transactions
- An initial visual exploration and analysis of the data set.

Classic Datasets: The graphics and visualization literature features datasets that have become classics, i.e., used as exemplars in hundreds of research papers. The first classic data set comes from computer graphics, namely, the teapot [Cro87]. Another classic data set is the Stanford Bunny [Sta93] originally published by Turk and Levoy [TL94]. The teapot and Stanford Bunny datasets are used as standard benchmarks for many rendering algorithms. Another classic data set that is used throughout the flow visualization literature is the tornado [CM93] release by Roger Crawfis. In addition, the Iris [Fis36] and Cars [Car05] data sets are featured in many parallel coordinate plots in the literature [War94]. We hope the dataset we describe here evolves into a classic financial transactions benchmark.

The rest of the paper is organized as follows: we present current barriers to bank transaction data in Section 1. In Section 2, we review the previous work on visual analysis of financial transactions

data. In Section 3, we describe the open retail bank transaction data set and categorization. Section 4 presents inital visualizations we created using the real bank data. Section 5 wraps up with conclusions and future work.

Barriers to Bank Transaction Data There are several barriers to accessing real bank transaction data, including data privacy and security concerns, regulatory restrictions, and the lack of standardization in data collection. Financial institutions cannot share sensitive information due to risk of breaches and violations of privacy regulations. Furthermore, many banks have their own systems for storing and processing transaction data, making it difficult to obtain a view of a financial history. Furthermore, there are bank transaction data sets available online that are not accessible directly or without an paid subscription or costly access fee.

2. Related Work

Overviews and Surveys: The book by Brose *et al.* [BFKN14] provides an overview of the visual analytics use in the field of finance and discuss components of visual analysis and its use in financial risk analysis. A survey by Ko *et al.* [KCA*16] focuses on approaches to visual analysis for exploring financial data. They classify financial systems according to data sources, applied automated techniques, visualization strategies, interactivity, and evaluation methods. The survey by Shi *et al.* [SLT*20] provides an overview of research that visually analyzes anomalous user behavior and categorizes them under the financial transaction domain that refers to money flows in buying and selling, as recorded in system logs. Another survey by Roberts and Laramee [RL18] highlights trends in business data visualization and visual analytic literature where visual analysis is utilized to address challenges stemming from business data, as well as industries that use visual design to expand their understanding of the business environment. The classification of literature covers subjects such as business intelligence, business ecosystems, and customer-centric data.

Visualization of Financial Transactions: The paper by Chang *et al.* [CGK*07] introduces WireVis, a multiview technique that helps analysts explore a large amount of categorical, time-varying data incorporating wire transactions. The work aims solving the problem of monitoring wire transactions in cooperation with Bank of America. The approach combines a search-by-example tool, a heatmap, a keyword network view, and a new visual design called Strings and Beads. All four views provide the user with a comprehensive representation of the links between the accounts, time, and keywords inside the transactions. Following this, Chang *et al.* [CLG*08] provide an overview of transaction data in the WireVis tool using a commercial relational database, while demonstrating that researchers can detect accounts and transactions that exhibit suspicious behavior. Joeng *et al.* [JDL*08] provide an exploratory user study to understand the relationship between user interaction and visual analysis, as well as an approach for capturing and evaluating user interactions while using the WireVis tool. The research by Arleo *et al.* [ATL*23] discusses the challenges of modelling financial dynamics and the need for a holistic understanding of the financial landscape. A visual analytics approach, Sabrina 2.0 is introduced, that supports exploration of financial data across different scales and generates firm-to-firm financial transaction networks to provide insight into the state of the economy.

Didimo *et al.* [DLM14] introduce VisFAN, a software tool for visualizing financial activity networks for crime detection. The tool features clustering algorithms and adjustable layout constraints management. They merge enhanced graph drawing methods with tools for social network analysis and automatic report generation to develop novel algorithms and interaction for visual analysis of networked datasets. The paper by Singh and Best [SB19] focuses on financial crime prevention by investigating and proving the use of visualization tools to aid in the detection of money laundering behavior trends. To investigate visualization techniques for identifying suspicious money transactions, a prototype, AML2ink, was created. The goal is to give an investigator a set of planned tests or analyses that visualize a group of transactions.

Leite *et al.* [LGM*17] provide a visual analytic system, EVA, a visual analytics approach for supporting financial fraud defections. Later, the same team [ALGM*20] proposes NEVA, the system is used for detection and analysis of fraudulent networks of bank transaction events. The system also enables exploring complex relations and dependencies of the data. Similarly, Maçãs *et al.* [MPM20] at introduce a visualization tool for analyzing banking transactions over time and detecting transaction topology and suspicious behavior. The work focuses on anonymized banking data provided by Feedzai, a fraud detection company, to develop a visual analytics tool for their analysts. A visual analytics tool, FinVis [RSE09] is developed to help the non-expert user to interpret the correlation aspects of financial data and make personal finance decisions while enabling them to assess potential long-term effects of various choices. Research by Xie *et al.* [XCH*14] introduces a visual analytics system, Visual Analysis of E-transaction Time-Series (VAET) that enables analysts to determine the key transactions in a vast dataset. The system enables users to analyze activities and provide a detailed view using a novel visual metaphor called KnotLines, where lines highlight the links between transactions and temporal trends.

Table 1 provides an overview of the related literature on money data, including a description of the data, its availability status (public or non-public), and information on where it can be accessed. Reasons for its restricted access are also provided. Transaction data is briefly described in the literature however it is not publicly available for privacy and security reasons. This is the inspiration for the work presented here.

3. MoneyData

The transaction data set spans 7 years starting in July of 2015. It contains over 6,500 retail bank transactions. Each transaction record features:

- Transaction date,
- Transaction type,
- Transaction description,
- Debit or credit amount,
- Remaining account balance.

Transaction type is a descriptor added automatically by the bank. We discuss this automatic categorization in Section 3.1.

Literature	Description of Data	Publicly Available	Where/Why not?
Chang *et al.* [CGK*07]	Financial transaction data provided by Bank of America	No	Privacy and proprietary reasons
Chang *et al.* [CLG*08]	Financial transaction data provided by Bank of America	No	Privacy and proprietary reasons
Joeng *et al.* [JDL*08]	Synthetic dataset contains 300 financial transactions involving 180 accounts with sender and receiver's names, date, keywords	No	Not shared
Didimo *et al.* [DLM14]	Example application data published by the Financial Crimes Enforcement Network	No longer	https://www.fincen.gov
Singh and Best [SB19]	Target branch and account, destination branch and account, cash flows, and the sum of amounts	No	Privacy and proprietary reasons
Leite *et al.* [LGM*17]	Money transactions data provided by the collaborating bank which contains 413 different accounts with 1,128,147 transactions and dimensions like sender/receiver,amount of money, location, and time of execution	No	Security and privacy reasons
Leite *et al.* [ALGM*20]	Money transactions data provided by the collaborating bank which contains 413 different accounts with 1,128,147 transactions and dimensions like sender/receiver,amount of money, location, and time of execution	No	Security and privacy reasons
Maçãs *et al.* [MPM20]	Anonymised transaction data provided by Feedzai, fraud detection company. The data contains client IBAN, location, amount, transaction, and date	No	Security and privacy reasons
Rudolph *et al.* [RSE09]	Not available	No	Not shared
Xie *et al.* [XCH*14]	Customer-to customer online retail business data which contains 26 million online e-transactions. About 9.3 million sellers and buyers are involved in the dataset.	No	Not shared

Table 1: The table provides an overview of the related literature on financial data, including a description of the data, its availability status (public or non-public), and information on where it can be accessed or reasons of the restriction.

Data Idiosyncrasies: Despite the transaction data being provided by a major retail bank, it does have some idiosyncrasies. One idiosyncrasy is the absence of transactions on weekends. Transactions that occur on Saturday or Sunday are archived as transactions on the following Monday, due to most UK banks implementing batch processing during business hours. Additionally, transaction data lacks timestamps, and although we generally believe transactions should appear in chronological order, there may be exceptions when manual processing or multiple parties are involved.

Anonymization In order to anonymize the data set, all identifier information is removed. This includes: account name, account number, sort code, and all other names of individuals in the transactions. All original names have been removed and replaced with pseudo-names.

3.1. Categorization

Given the set of transactions, we attempted some hierarchical categorizations. The first two categories of transactions are high-level: Credit and Debit. Then we discussed various ways to add another layer to the hierarchy of categories: manual, semi-automatic, and automatic.

Manual Categorization We made an attempt to add descriptive categories to each transaction by manual inspection of the data records. By putting the data into a Google Sheet, we can sort the record such that tuples with identical description will be grouped together to facilitate the categorization. We added categories such as Travel, Supplementary Income, Services, Savings, Paycheck, Shopping, Mortgage Payment, Investments, Interest, Home Improvement, Health, Groceries, Fitness, Entertainment, Dining Out, Clothing, Cash, Utility Bills, Amazon. By adding these categories we can create hierarchical visual representations. We can add another level of detail by depicting all of the transactions inside an individual category. We note that this categorization is arbitrary and other categorizations can be used.

We also made an attempt to add location information to the transactions, however, this kind of meta-data contains more uncertainty than the other categories. For example a Google query can be made using the description field of the transaction, sometimes revealing the city the transaction took place in. This attribute is also incomplete. Users that are not interested in this uncertain data attribute may delete it.

Semi-automatic Categorization We have created a python script that takes the data as an argument to categorize transactions in the input into different categories and sub-categories based on their transaction types and descriptions. The script searches for specific key words and phrases and assigns a sub-category based on those key words. It then creates a new column named category spend, which assigns a category to each transaction. It assigns different categories such as bill payments, cash points, account fees, transfers, check payments, deposits, income, shopping, and others based on various transaction types.

The script then creates a new column named sub-category, which assigns a sub-category to each transaction based on the type and description. It assigns various sub-categories such as in-store purchase debit, online shopping debit, cash point withdrawals, bill payments, savings, money transfer debit, and bank fee credit, among others, based on specific criteria. This provides a systematic approach to categorizing transactions in the given data into specific categories and sub-categories, which can help analyze and visualize spending patterns or track expenses in financial data. The python script can be downloaded from GitHub at: **https://github.com/thevisgroup/MoneyVis**.

Automatic Categorization The bank from which the transactions are archived provide a categorization, or transaction type. The transaction types and associated labels are as follows (Code, description):

- BGC: Bank Giro Credit, BP: Bill Payment
- C/P: Cashpoint, CHQ: Cheque
- D/D: Direct Debit, DEB: Payment type Debit Card
- DEP: Deposit, FEE: Fixed Service Charge
- FPI: Faster Payments Inwards, FPO: Faster Payments Outwards
- PAY: Payment, TFR: Transfer, SO: Standing Order

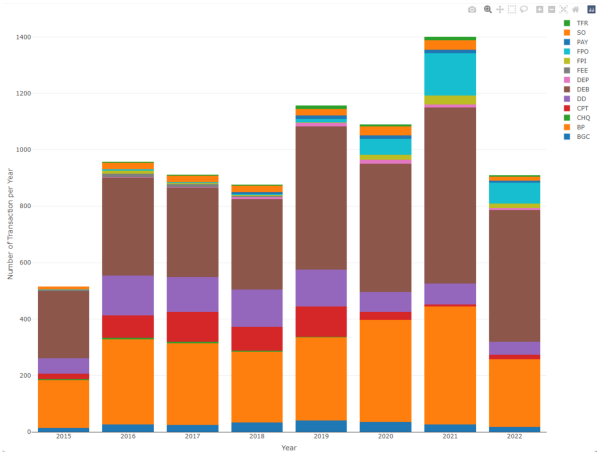

Figure 1: This image shows the number of transactions (y-axis) and bank transaction categories in each year from 2015-2022. Bank transaction codes are provided in Section 3.1.

3.2. Data Access

The following URL provides read access to the open bank transaction data set:
https://tinyurl.com/y4e8yevn

4. Initial Visualizations

In this section, we present a series of initial images created using the money data we introduce (see Section 3). These images showcase preliminary characteristics of the data and incorporate our classification system, enabling understanding of some patterns. They offer a starting point.

Figure 1 presents a stacked bar chart. The chart focuses on the spending from 2015–2022 with the total amount spent in each category. This image provides a more detailed view of the trends in transaction volume over the years. The bank transaction code and description provided by the bank can be found in Section 3.1, providing further context and understanding of the different transaction categories. This figure shows the main transactions were made for standing orders (SO), debit card payments (DEB), and Bill Payments (BP). We can observe a significant increase in spending in 2021 especially with the faster payments outwards (FPO) and observe the last check written in 2019 (CHQ).

Figure 2 shows locations of transactions with pie charts placed where spending occurred. The top one displays the amount of money spent in each category (see Section 3.1) within different cities in the UK. This image provides an understanding of how spending patterns vary across different regions in the UK, enabling visual analysis of regional spending preferences. Home improvement, groceries, and Amazon payments appear to be the main expenses among all categories in a given region.

The figure 3 shows a sunburst chart divided into segments based on the number of transactions: outgoing, incoming, and savings. The figure displays the total incoming categories and outgoing transactions. This enables for an analysis of how the incoming money was spent in each category. The image displays that investment is the main investment category accounting for 13% of the total transactions.

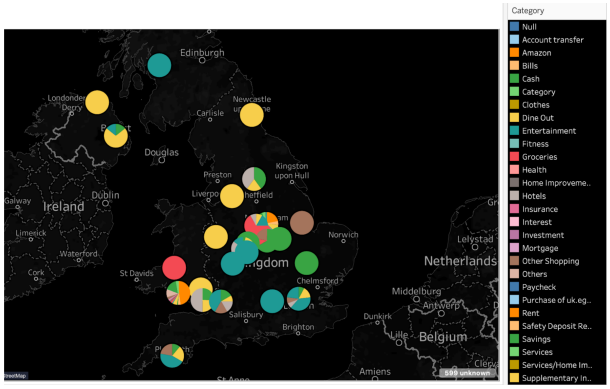

Figure 2: (Top) The figure shows money spent in each category mapped to location in the UK only. (Bottom) The figure reveals a closer view of the money spent in some of the cities.

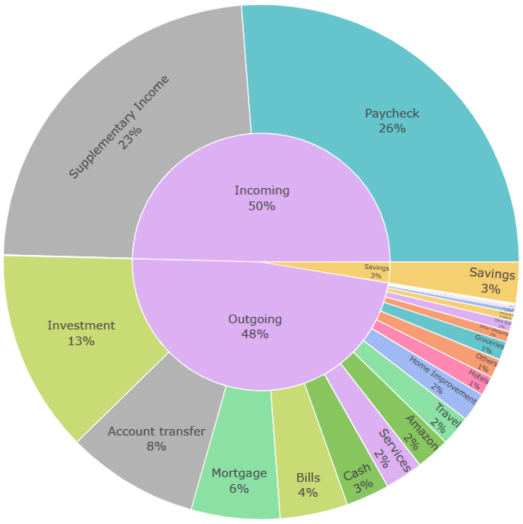

Figure 3: The sunburst chart displays the total incoming and outgoing transactions. The segments are based on the total amount rather than the number of transactions. The colors here are arbitrary.

We offer a supplementary video with this submission that shows the interaction with the images presented here.

5. Conclusions and Future Work

To our knowledge, we present the first open, retail, bank transaction data set for visualization and analysis purposes. We believe this will be a valuable asset when developing finance-based visual analysis tools and other software that processes similar transactions, e.g., fraud detection. As such, we believe this offers a plethora of future work directions. For example, we would like to apply machine learning for automatic categorization of transactions and predictive analysis. We would also like to apply a third party annotation system such as Open Works Annotation to the transactions. The transaction data can form the basis of many case studies since it is historic. For example the analysis of salary versus inflation over time, the study of cost-of-living, the study of spending habits and routines, risk-level assessment for loans, the effectiveness of programs such as ”save the change”, extracting periodic behavior and so on. And of course the open bank transaction data set can be used

as a benchmark for popular existing visual designs and software and educational purposes.

6. Acknowledgements

The project was supported, in part, by EPSRC funding from EP/S010238/2.

References

[ALGM*20] A. LEITE R., GSCHWANDTNER T., MIKSCH S., GSTREIN E., KUNTNER J.: Neva: Visual analytics to identify fraudulent networks. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 344–359. 2, 3

[ATL*23] ARLEO A., TSIGKANOS C., LEITE R. A., DUSTDAR S., MIKSCH S., SORGER J.: Visual exploration of financial data with incremental domain knowledge. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, pp. 101–116. 2

[BFKN14] BROSE M. S., FLOOD M. D., KRISHNA D., NICHOLS B.: *Handbook of Financial Data and Risk Information II*, vol. 2. Cambridge University Press, 2014. 2

[Car05] Car dataset, 2005. Last accessed: February 2023, http://lib.stat.cmu.edu/datasets/cars.data. 1

[CGK*07] CHANG R., GHONIEM M., KOSARA R., RIBARSKY W., YANG J., SUMA E., ZIEMKIEWICZ C., KERN D., SUDJIANTO A.: Wirevis: Visualization of categorical, time-varying data from financial transactions. In *2007 IEEE symposium on visual analytics science and technology* (2007), IEEE, pp. 155–162. 2, 3

[CLG*08] CHANG R., LEE A., GHONIEM M., KOSARA R., RIBARSKY W., YANG J., SUMA E., ZIEMKIEWICZ C., KERN D., SUDJIANTO A.: Scalable and interactive visual analysis of financial wire transactions for fraud detection. *Information visualization 7*, 1 (2008), 63–76. 2, 3

[CM93] CRAWFIS R. A., MAX N.: Texture splats for 3d scalar and vector field visualization. In *Proceedings Visualization'93* (1993), IEEE, pp. 261–266. 1

[Cro87] CROW F.: The origins of the teapot. *IEEE Computer Graphics and Applications 7*, 1 (1987), 8–19. 1

[DLM14] DIDIMO W., LIOTTA G., MONTECCHIANI F.: Network visualization for financial crime detection. *Journal of Visual Languages & Computing 25*, 4 (2014), 433–451. 2, 3

[Fis36] FISHER R. A.: The use of multiple measurements in taxonomic problems. *Annals of eugenics 7*, 2 (1936), 179–188. 1

[JDL*08] JEONG D. H., DOU W., LIPFORD H. R., STUKES F., CHANG R., RIBARSKY W.: Evaluating the relationship between user interaction and financial visual analysis. In *2008 IEEE Symposium on Visual Analytics Science and Technology* (2008), IEEE, pp. 83–90. 2, 3

[KCA*16] KO S., CHO I., AFZAL S., YAU C., CHAE J., MALIK A., BECK K., JANG Y., RIBARSKY W., EBERT D. S.: A survey on visual analysis approaches for financial data. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 599–617. 2

[LGM*17] LEITE R. A., GSCHWANDTNER T., MIKSCH S., KRIGLSTEIN S., POHL M., GSTREIN E., KUNTNER J.: Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics 24*, 1 (2017), 330–339. 2, 3

[MPM20] MAÇÃS C., POLISCIUC E., MACHADO P.: Vabank: visual analytics for banking transactions. In *2020 24th International Conference Information Visualisation (IV)* (2020), IEEE, pp. 336–343. 2, 3

[RL18] ROBERTS R. C., LARAMEE R. S.: Visualising business data: A survey. *Information 9*, 11 (2018), 285. 2

[RSE09] RUDOLPH S., SAVIKHIN A., EBERT D. S.: Finvis: Applied visual analytics for personal financial planning. In *2009 IEEE symposium on visual analytics science and technology* (2009), IEEE, pp. 195–202. 2, 3

[SB19] SINGH K., BEST P.: Anti-money laundering: using data visualization to identify suspicious activity. *International Journal of Accounting Information Systems 34* (2019), 100418. 2, 3

[SLT*20] SHI Y., LIU Y., TONG H., HE J., YAN G., CAO N.: Visual analytics of anomalous user behaviors: A survey. *IEEE Transactions on Big Data* (2020). 2

[Sta93] Stanford bunny, 1993. Last accessed: February 2023, https://faculty.cc.gatech.edu/~turk/bunny/bunny.html. 1

[TL94] TURK G., LEVOY M.: Zippered polygon meshes from range images. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques* (1994), pp. 311–318. 1

[War94] WARD M. O.: Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings Visualization'94* (1994), IEEE, pp. 326–333. 1

[XCH*14] XIE C., CHEN W., HUANG X., HU Y., BARLOWE S., YANG J.: Vaet: A visual analytics approach for e-transactions time-series. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 1743–1752. 2, 3