data.org

TECHNICAL CASE STUDY

Privacy-Enhancing Technologies (PETs) for Public Health

Purpose

Motivated by its mission to accelerate the social impact sector's ability to optimally leverage data and technology, data.org partnered with Mastercard, Harvard University, OpenDP, Javeriana University, and Sloan Foundation, to run a unique Privacy-Enhancing Technologies (PETs) for Public Health Challenge to unlock innovation in the use of financial transaction data, combined with cross-domain datasets, to develop decision support tools and solutions.
This technical case study presents a detailed analysis of how the combination of the PETs technology, partnership ecosystem, derivative datasets, and analytical and evaluative processes were successfully applied to execute the novel PETs for Public Health Challenge. It also covers several of the challenges faced, solutions implemented, and the project outcomes.

It is envisioned that this technical case study will help professionals and businesses across private sector, research and academic institutions, and social impact organizations, etc. to understand best practices, learn from our experiences, and where applicable apply similar strategies to their own situations to advance data sharing for social impact.

Acknowledgment

This technical case study report would not be possible without the contributions of the experts and ecosystem actors who shared their expertise at various stages of the project. We are grateful for the time and effort they have given to this project. To the Project Team and data.org colleagues, especially Avinash Laddha, who pushed this work forward, interpreted insights, and contributed to this final report: thank you!

- **Mastercard:** Shanna Crumley, Smita Jain, John Derrico, Chapin Flynn
- **data.org:** Uyi Stewart Ph.D., Danil Mikhailov Ph.D., Avinash Laddha, Paul Korir Ph.D., Hugo Gruson Ph.D., David Mascarina, Stephanie Earl.
- **Harvard University/OpenDP:** Salil Vadhan Ph.D., Sharon Ayalde, Yanis Vandecasteele, Vikrant Singhal,
- **University of Javeriana:** Zulma Cucunuba Ph.D., Nicolas Dominguez, Felipe Segundo, Diana Fajardo,
- **Judges**: Bubacarr Bah, Ph.D., Sean Cavany, Ph.D., John Derrico, Jack Fitzsimons, Chapin Flynn, Christine Task, Ph.D., Charlie Whittaker, Ph.D., Wanrong Zhen, Ph.D.
- **Global Challenge Awardees:** Kris Parag Ph.D., (Imperial College London, UK); Ben Lambert Ph.D., (University of Oxford, UK); Anil Vullikanti Ph.D., Zihan Guan, and Dung Nguyen (University of Virginia); B. Aditya Prakash, and Leo Zhao (Georgia Tech); Ravi Tandon, Payel Bhattacharjee, and Fengwei Tian (University of Arizona); Dmitrii Usynin (Technical University of Munich, Germany); Shubham Kumar, Milan Anand Raj, and Divya Gupta (Indian Institute of Technology Kanpur, India)

Finally, thank you also to the Sloan Foundation for their financial support to run the challenge and implement this project.

Uyi Stewart, Ph.D.,
Chief Data & Technology Officer, data.org

Foreword

Back in the Fall of 2020, as I was doing my final interviews for the role of Executive Director of an exciting new nonprofit, founded by the Mastercard Center for Inclusive Growth and Rockefeller Foundation, called data.org, Mike Froman, the then Vice Chairman and President of Strategic Growth at Mastercard, posed me an unexpected challenge: if I got this role, how would I go beyond only looking to Mastercard for funding, but also partner with them to leverage data and their data science talent for social impact? Thinking on my feet, I ended up pitching the project that ended up being the Privacy-Enhancing Technologies (PETs) for Public Health Challenge. The reality, of course, is much more complex than such a simple origin story might suggest. In my previous role as Head of Data & Innovation at the Wellcome Trust, a global funder of health programs, I was already thinking about how new data sources outside of health data could help modellers and public health officials better understand the link between human behaviour, like purchasing and mobility, and the spread of infectious disease. My good friend and colleague, Dr Uyi Stewart, was doing the same at the same time at the Gates Foundation, and so were many others, including Mastercard itself.

All of us understood that technology companies like Mastercard were custodians of data that would be immensely useful for research and policy, because of their global reach, scale, and local penetration, with the products and solutions they provide being used daily by billions of people. If only these organizations could be persuaded that there were technologies and processes now in place to use such data safely, respecting both privacy of their customers and the commercial sensitivity of the corporations themselves. Thus, the PETs for Public Health Challenge was born.

Danil Mikhailov, Ph.D., Executive Director, data.org

At Mastercard, we believe when it comes to the data of our customers, partners, or individual cardholders, You own it. You control it. You should benefit from the use of it. We protect it. Those statements are at the core of our business, so when we started thinking about how to use data to make a positive impact on communities and individuals around the world, we knew we had to do it the right way. We had to be sure we were adhering to all those principles while finding ways to leverage data for social impact and inspiring other private companies to do the same. So, the question arose… how could we confidently achieve all those goals? Turns out the answer to that question was to connect some of the smartest people in the world, who focus on privacy, encryption, data governance, and data science and epidemiological modelling, with some of the latest privacy-enhancing technologies, like differential privacy. The result was the amazing piece of work that you have in front of you. We sincerely hope this work inspires other private sector companies to consider the ways in which data can be used, in a safe, responsible, and transparent manner, to drive meaningful, lasting social impact.

Chapin Flynn, Senior Vice President, Transit and Urban Mobility, Mastercard

Executive Summary

Mastercard, a global technology company, sought to explore how data may be used in a privacy preserving environment to support, ***at scale***, researchers who are working to advance social impact causes, e.g., public health pandemic management. By applying Differential Privacy (DP) to synthetic transaction data, data.org and collaborators (Harvard University, OpenDP, and Javeriana University) successfully ran a PETs for Public Health Challenge that created six reusable tools/frameworks (see descriptions in section 4) to support data-driven insights for pandemic management. Through this process, we developed an approach on how to generate realistic synthetic transaction data.

Our work has shown that synthetic transactional data combined with mobility and other public health datasets hold significant spatial-temporal information, which can reveal real-time behavioural patterns of populations. When analysed through a differential privacy lens, we found that such data have useful predictive power - actionable insights - for so-called nowcasting (immediate hotspot detection, mobility patterns), and forecasting (identifying future infection rates and contact matrixes) summarized in the following three areas:

1. **Diagnostic support for Nowcasting, i.e., the ability to explain the present or very near future.** Two scenarios to note:
    a. Hotspot Detection: Identifying areas of high physical interaction to prioritize resource allocation.
    b. Pandemic Adherence Monitoring: Tracking shifts in spending behaviour in response to lockdowns and other interventions.
2. **Predictive capacity for forecasting, i.e., predicting future events, trends, or outcomes.** Two scenarios to note:
    a. Mobility Analysis: Understanding movement patterns to guide quarantine policies and predict infection trends.
    b. Contact Matrix Estimation: Estimating interaction frequencies across age groups to assess transmission risks.
3. **Development of learning models, e.g., epidemic management.**
    a. We developed 2 ML/AI frameworks with the ability to incorporate financial transaction data along with other diverse datasets to support the scenarios outlined in 1 and 2 above, as well as relevant new use cases.

Introduction

Over the past decade, fueled by advances in Machine Learning (ML), Artificial Intelligence (AI), and other emerging technologies, the use of ***private sector data***, such as mobile phone data (integrated with other cross-domain datasets) **to develop new business** and social impact **solutions** has been steadily growing. For example, it has been used to optimize efficiencies in the gig economy with Uber, Lyft, to develop a bustling FinTech market in Nigeria, and to manage endemic diseases linked to the seasonal patterns of dengue in Pakistan and rubella in Kenya. [1]

However, there is now a heightened awareness of the risk associated with poor data management, resulting in an escalation in the strictness of laws and regulations that stipulate how data may be handled and combined. Consequently, the advent of data protection regulation in over 71% of countries around the world, with a further 9% with draft legislation underway, has raised the urgency for technologies that can enable data integration and analytics while preserving privacy. [2] Privacy-Enhancing Technologies (PETs) have emerged as a promising solution to navigate this dilemma, enabling secure data integration and analytics for business and social impact applications while preserving individual privacy. Newer PETs methods, such as Differential Privacy (DP), have been shown to provide statistically verifiable protection against identifiability.

While PETs have been around for a while, they have largely been a frontier technology that only a select number of regulators and a few companies in the private sector have had an interest in exploring and the luxury of doing so.

[1]https://link.springer.com/article/10.1186/s40537-021-00553-4#:~:text=The%20main%20contribution,data%20processing%20formatHYPERLINK "https://www.federalreserve.gov/econres/notes/feds-notes/electricity-demand-as-a-high-frequency-economic-indicator-20201021.html"https://www.federalreserve.gov/econres/notes/feds-notes/electricity-demand-as-a-high-frequency-economic-indicator-20201021.htmlHYPERLINK "https://bmjopen.bmj.com/content/14/11/e083096"https://bmjopen.bmj.com/content/14/11/e083096

[2] https://unctad.org/page/data-protection-and-privacy-legislation-worldwide?utm_source=chatgpt.com

Problem Statement

Pandemics are an unavoidable global threat that necessitate rapid and informed public health responses. Effective management requires granular, diverse, and timely data to inform public health strategies such as tracking disease spread, mapping mobility patterns, and managing human behavioural changes. However, getting such granular, diverse, and timely data remains a major hurdle. For example, research has shown that traditional data collection methods have limitations and potential pitfalls due to the challenges of combining disparate data sources [1], [2], [3], [4]. To this end, the aims of this project were twofold:

1. Assess the feasibility of applying PETS (such as DP) to synthetic transaction data with a view of informing optimal epidemiological decision-making while preserving privacy in the underlying data, and
2. Collaboratively develop generalizable methods for integrating cross-domain data sets to support practical applications of the PETs framework applied to other use cases.

To achieve these goals, we had to address the following technical challenges:

1. **Data Management & Privacy:** How to develop realistic synthetic datasets with a baseline privacy-preserving template
2. **Translational Use Cases:** How to generate generalizable use cases for pandemic management
3. **Application:** How to apply the use cases to develop applications or solutions from the interpolated & privacy-preserved datasets.

Here's a closer look at each challenge and how we approached them:

1. How to develop synthetic dataset with a baseline privacy-preserving template?

Given the inherent sensitivity of financial data, the creation of privacy-preserved synthetic datasets becomes a crucial alternative. We developed a synthetic financial dataset in collaboration with Harvard, incorporating a baseline privacy-preserving template.

The synthetic dataset creation process addressed the sensitive nature of the underlying data by creating a privacy-preserving template including:

- **Mimicking Real Data Format:** The synthetic data replicated the structure of actual transactions without requiring the disclosure of any financial records. Rather, the synthetic dataset was generated using a structured data dictionary designed to mimic realistic weekly merchant transaction behaviour across multiple cities and industries. Key variables included:
    - ID & Merchant ID: Unique identifiers ranging from 1 to 10,000.
    - Date: Captures end-of-week timestamps from January 1, 2019, to December 27, 2022, with weekly intervals.
    - Merchant Category: Represents industry types such as Airlines, Restaurants, Drug Stores/Pharmacies, etc., covering a broad range of consumer sectors.
    - Merchant Postal Code: Encodes merchant location, allowing city-level identification using postal code patterns — Medellin ('05'), Bogota ('11'), Brasilia ('-000'), and Santiago (default).
    - Transaction Type: Indicates whether the transaction occurred ONLINE or OFFLINE, reflecting the sales channel.
    - Spend Amount (spendamt): The total weekly transaction value per merchant.
    - Number of Transactions (nb transactions): The count of individual transactions per merchant per week.
- **Integrating Public Real-World Information:** Postal codes aligned with specific city subdivisions (Medellin, Bogota, Brasilia, Santiago), and merchant categories were inspired by Mastercard's publicly available guidelines (https://www.mastercard.us/content/dam/public/mastercardcom/na/global-site/documents/quick-reference-booklet-merchant.pdf). Crucially, real-world COVID-19 data from "Our

World in Data" informed the modelling of financial responses to pandemic trends, allowing for realistic simulation without exposing private health or financial details.

- **Leveraging Bayesian Prior Knowledge (general patterns):** We leveraged domain expertise to enhance the realism of the data by calibrating merchant category frequencies (e.g., by applying the mechanism to use #death as a factor in average spend amount, we also tried to simulate realistic changes in earnings or spending behavior based on how many new deaths are happening). We also leveraged Mastercard's reference guide (above) to calibrate Merchant category frequencies to reflect typical industries' distribution within the cities. In addition, we applied a COVID-19 effect multiplier to simulate the pandemic's impact on spending – essentially this means we were able to manage some expected/random penalties (Expected drop in count/sum of amount spent) applied to each merchant category.
- **Applying a Baseline Privacy Algorithm:** Recognizing the potential for even synthetic data to inadvertently reveal patterns related to the real data it was based on, a privacy-preserving algorithm, from OpenDP library, was implemented. This crucial step ensured that the shared dataset offered a measurable level of privacy protection, further mitigating any risk of inference about the original information.

The resulting privacy-preserved synthetic dataset allowed teams to develop and test their epidemic analysis solutions in a safe environment, directly addressing the fact that the underlying financial data would otherwise be completely out of reach for such a challenge. This approach successfully facilitated innovation in data analysis for public health while firmly respecting and upholding data privacy principles.

2. How to generate generalizable use cases for pandemic management?

Once we created the baseline structure for the dataset, the biggest challenge we encountered was how to generate real-world pandemic use cases to facilitate the use of the dataset for the development of applicable decision support tools. We addressed this challenge in two steps: first, the creation of generalizable use cases, and second, adapting the synthetic dataset to fit the use cases. These are described below:

Step 1: Creation of Generalizable Use Cases:

We created an evaluation framework that was structured around five distinct policy use cases or scenarios based on generalizable data science underpinnings (nowcasting, forecasting, learning models, etc.). Each scenario represents a critical challenge in pandemic management where the integration of privacy-enhanced transactional data holds significant promise. These use cases or scenarios defined the specific technical objectives and expected outcomes for tools, solutions, and frameworks that were developed.

Use Case 1: Enhancement of Epidemiological Techniques through Privacy-Preserving Data Integration

- Technical Challenge: How can privacy-enhanced transactional data be seamlessly integrated with established epidemiological techniques to provide a more granular and timely understanding of disease transmission dynamics?
- Expected Outcome: Development of methodologies and tools that demonstrate the effective incorporation of financial data, secured through differential privacy, into standard epidemiological analyses, leading to improved public health response in real time. This includes the design of agile tools for joint analysis of financial and open datasets.

Use Case 2: Inferring Contact Patterns and Constructing the Who Acquires Infection from Whom (WAIFW) Matrix

- Technical Challenge: How can privacy-enhanced transactional data be utilized to derive detailed insights into contact patterns across diverse population segments (e.g., age groups, professions, commercial activity involvement)?

- Expected Outcome: Development of techniques that leverage transactional data to inform the who acquires infection from whom (WAIFW) matrix, providing a more accurate representation of infection transmission pathways and enabling more targeted public health interventions.

Use Case 3: Real-time Estimation of the Effective Reproduction Number (Rt)

- Technical Challenge: How can privacy-enhanced transactional data, specifically in the context of informing contact patterns, contribute to more accurate and timely real-time Rt estimations?

- Expected Outcome: Development of models and algorithms that integrate transactional data-derived contact information to enhance the precision and responsiveness of Rt calculations, a key metric for assessing the trajectory of an epidemic.

Use Case 4: Mitigation of Bias in Nowcasting Estimations due to Behavioural Factors

- Technical Challenge: How can privacy-enhanced transactional data be employed to identify and correct inherent biases in nowcasting estimations that arise from population behavioural changes and reporting lags?

- Expected Outcome: Development of statistical methodologies that leverage transactional data as an indicator of behavioural shifts, enabling more accurate real-time predictions and improving overall situational awareness for public health officials.

Use Case 5: Integration of Transactional Data as a Predictive Feature in Epidemic Forecasting

- Technical Challenge: How can privacy-enhanced transactional data be effectively incorporated as a predictive variable in models designed for forecasting epidemic curves (e.g., cases, deaths, Rt), with the goal of enhancing predictive accuracy?

- Expected Outcome: Development and evaluation of forecasting models that demonstrate the added value of transactional data in improving the reliability and robustness of short-term and potentially long-term epidemic projections.

Step 2: Adapting the Synthetic Dataset to the Generalizable Use Cases:

We designed specific features in the synthetic dataset to replicate some expected features of the financial time series. We developed Python code to generate a mock dataset based on the data dictionary provided by Mastercard that simulated the financial time series for each merchant category based on ad-hoc parameters correlating them with epidemiological time series extracted from openly available datasets. To ensure consistency with the available epidemiological data and expected trends from each merchant category, we implemented the correlation analysis to refine the data generation process, ensuring better alignment with observed phenomena during the COVID-19 pandemic, particularly regarding correlations with epidemiological metrics.

1. Recommendations for Epidemiological Metrics

Switching from New Cases to New Deaths: Initially, the mock dataset was generated using the incidence of cases as the epidemiological metric to correlate financial transactions with pandemic trends. **We recommended using the incidence of deaths instead of the incidence of cases.** The reason for this is that new cases can be highly influenced by external factors, such as testing availability and transmissibility of variants. For instance, during the Omicron wave, the high transmissibility resulted in significantly more cases but not a proportional increase in deaths. Death counts are often more reflective of the pandemic's impact on policy decisions and public behaviour, which influences transactional trends. Using deaths helped mitigate unintended drastic shifts in transactional patterns during periods with inflated case numbers.

2. Adjustment of Transaction Sampling Proportions

City-Specific Proportions: We suggested that the number of transactions in each city should reflect the proportion of its population relative to the total population in the dataset. The following population percentages were recommended for scaling:

```
self.populations = {"Medellin": 2569000,
                    "Bogota DC": 7181000,
                    "Brasilia": 4935000,
                    "Santiago": 5561000}
total_population = sum(self.populations.values())
percentages = {key: value/total_population*100 for key, value in
self.populations.items()}
```

```
{'Medellin': 12.688926207645954,
 'Bogota DC': 35.46873456485232,
 'Brasilia': 24.375185221772202,
 'Santiago': 27.467154005729526}
```

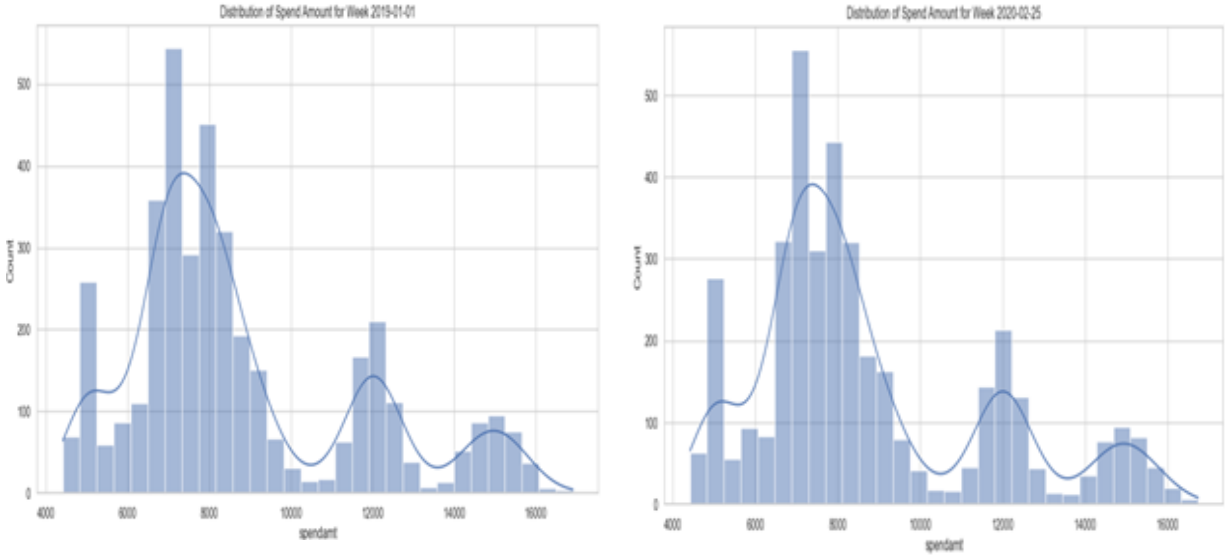
This adjustment ensured aligning the synthetic dataset aligned with realistic demographic distributions.

3. Merchant Category Spending Trends

Negative observations in the number of transactions: An initial review of the mock dataset showed the presence of some negative numbers of transactions, which were replaced by 1 to run the correlation analysis; meanwhile, OpenDP's representative resolved the issue by adjusting the number of transactions.

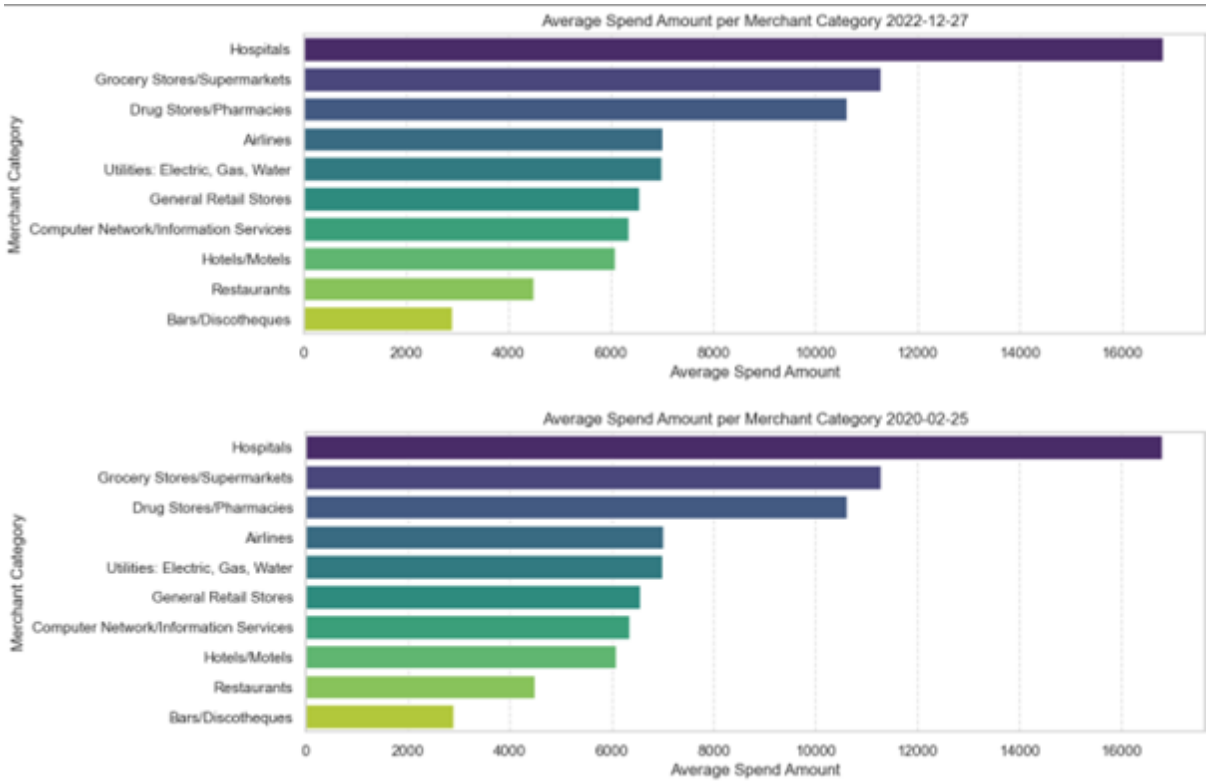
```
> df_mock %>% group_by(nb_transactions) %>% count()
# A tibble: 378 x 2
# Groups:   nb_transactions [378]
   nb_transactions     n
             <int> <int>
 1              -5     1
 2              -3     1
 3              -2     4
 4              -1     5
 5               0    62
 6               1   112
 7               2   260
 8               3   498
 9               4  1043
10               5  1989
# i 368 more rows
# i Use `print(n = ...)` to see more rows
```

Average Spending Amount: An analysis of the average spending per merchant category at different pandemic stages indicated minimal changes.



This stability seemed to have been due to the choice of multipliers for each category and the static nature of typical spending amounts. We suggested reviewing and adjusting these multipliers to better reflect dynamic changes during the pandemic.

Frequency Distribution: A similar issue was observed in the frequency distribution of spending amounts at the pandemic's onset, as exemplified by the following histograms, which correspond to different weeks (2020-02-25 and 2022-12-27, respectively).



While clustering effects emerged later, the **early-stage stability suggested the need to refine the modeling of spending dynamics during that period**, which was corrected by OpenDP's representative.

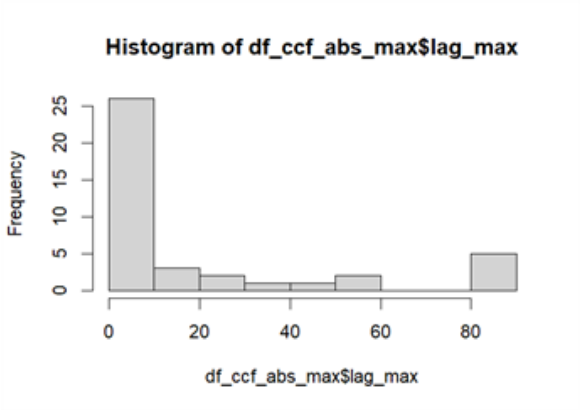
Correlation Analysis Results

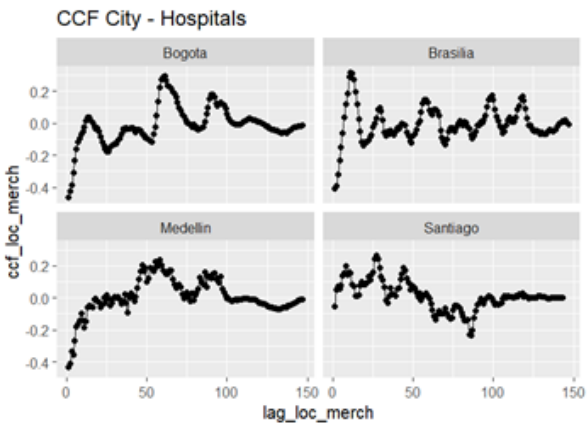
Positive Findings: The cross-correlation analysis showed promising results, with most lags clustering around 0–3 weeks, indicating a strong and timely relationship between epidemiological data and transactional patterns:

```
> df_ccf_abs_max %>% arrange(admin_name1) %>% print(n = 50)
# A tibble: 40 × 4
   merch_category                      admin_name1 lag_max ccf_max
   <chr>                               <chr>         <dbl>   <dbl>
 1 Airlines                            Bogota            1   0.745
 2 Bars/Discotheques                   Bogota            1   0.925
 3 Computer Network/Information Services Bogota          0  -0.431
 4 Drug Stores/Pharmacies              Bogota            1  -0.562
 5 General Retail Stores               Bogota            1   0.739
 6 Grocery Stores/Supermarkets         Bogota            2   0.371
 7 Hospitals                           Bogota            0  -0.476
 8 Hotels/Motels                       Bogota            2   0.757
 9 Restaurants                         Bogota            1   0.885
10 Utilities: Electric, Gas, Water     Bogota           34   0.149
11 Airlines                            Brasilia          2   0.752
12 Bars/Discotheques                   Brasilia          1   0.867
13 Computer Network/Information Services Brasilia        0  -0.205
14 Drug Stores/Pharmacies              Brasilia          2  -0.540
15 General Retail Stores               Brasilia          1   0.699
16 Grocery Stores/Supermarkets         Brasilia          3   0.449
17 Hospitals                           Brasilia          1  -0.410
18 Hotels/Motels                       Brasilia          1   0.720
19 Restaurants                         Brasilia          2   0.851
20 Utilities: Electric, Gas, Water     Brasilia         13  -0.206
21 Airlines                            Medellin          1   0.585
22 Bars/Discotheques                   Medellin          1   0.849
23 Computer Network/Information Services Medellin       55   0.259
24 Drug Stores/Pharmacies              Medellin          1  -0.505
25 General Retail Stores               Medellin          2   0.653
26 Grocery Stores/Supermarkets         Medellin          0   0.385
27 Hospitals                           Medellin          0  -0.481
28 Hotels/Motels                       Medellin          0   0.699
29 Restaurants                         Medellin          1   0.813
30 Utilities: Electric, Gas, Water     Medellin         21  -0.216
31 Airlines                            Santiago         86   0.323
32 Bars/Discotheques                   Santiago         86   0.445
33 Computer Network/Information Services Santiago       19  -0.169
34 Drug Stores/Pharmacies              Santiago         48  -0.209
35 General Retail Stores               Santiago         86   0.371
36 Grocery Stores/Supermarkets         Santiago         13  -0.250
37 Hospitals                           Santiago         27   0.263
38 Hotels/Motels                       Santiago         86   0.321
39 Restaurants                         Santiago         86   0.441
40 Utilities: Electric, Gas, Water     Santiago         53  -0.267
```

Anomalous Observations:

- Utilities (Electric, Gas, Water): These categories show unusually high lags (e.g., 34 weeks in Bogota), likely due to the assigned multipliers being 0, indicating weak or no correlation.
- Hospitals: Negative correlations are observed in 3 of the 4 cities (e.g., Bogota, Brasilia, and Medellin), despite positive multiplier





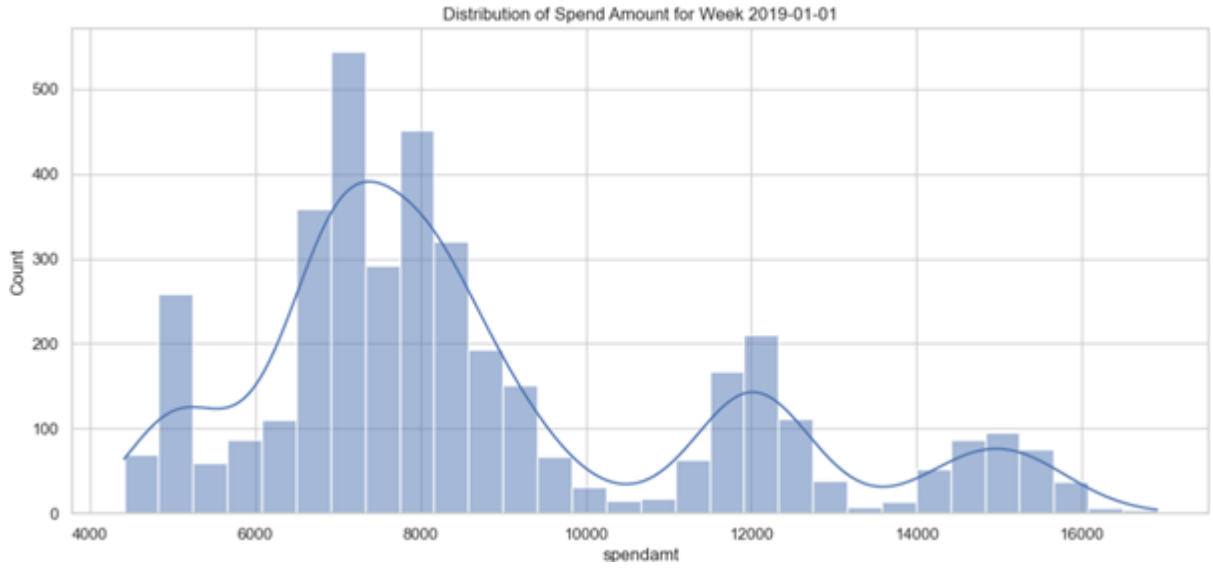
We suggested correcting this by switching from cases to deaths as the epidemiological metric.

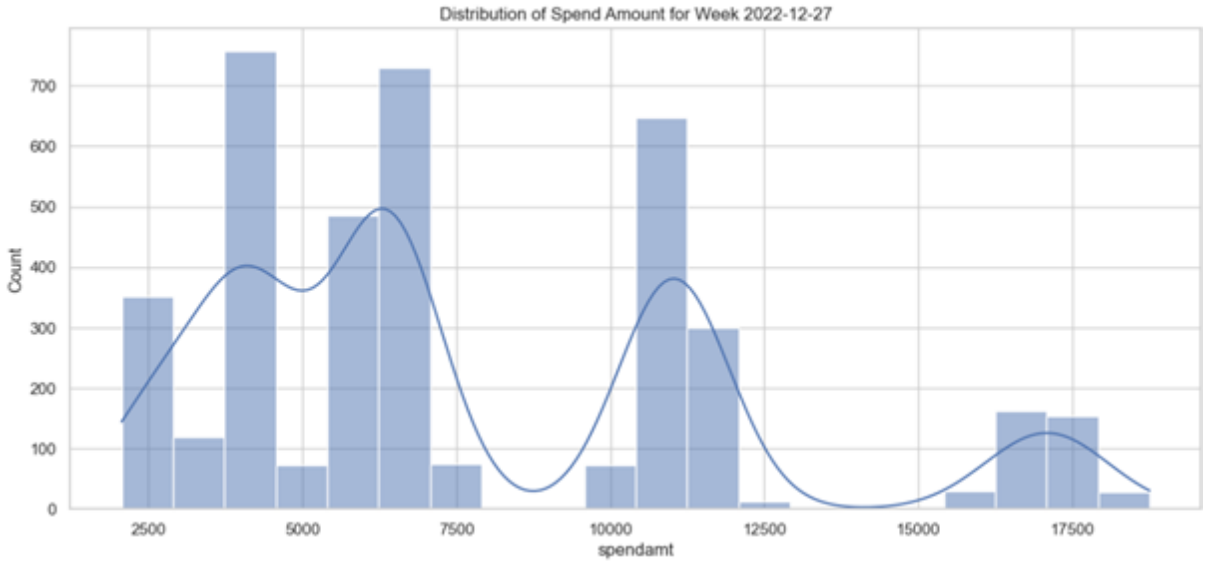
Summary:

To improve the dataset and its alignment with real-world dynamics, the following steps were recommended:

1. Replace new cases with new deaths as the epidemiological metric to correlate the financial mock dataset with pandemic trends.
2. Adjust transaction sampling proportions based on city populations.
3. Revisit and refine the spending multipliers for each merchant category to reflect dynamic changes during the pandemic.
4. Investigate and resolve anomalies in specific merchant categories, such as Utilities and Hospitals.

These adjustments helped enhance the synthetic dataset's realism and ensure it served as a reliable basis for analysing the interplay between transactional and epidemiological data.





3. How to apply the use cases to develop applications or solutions from the interpolated & privacy-preserved datasets

Through the challenge, we developed a suite of tools leveraging Privacy-Enhancing Technologies (PETs) to analyze sensitive data to address each of the five use cases outlined without compromising individual privacy.