The Value of Sharing Market Data through Data Analytics on Digital Platforms: Empower Small Businesses on Alibaba’s Taobao Marketplace

Abstract

Despite the promising value of data in the digital markets, data sharing across firms is hindered due to increasingly stringent data privacy and security regulations. Data analytics enable firms to extract value from data while mitigating privacy concerns from directly sharing data. Leveraging their remarkable data aggregation and analytical capabilities, digital platforms can play an essential role in empowering firms, particularly small businesses, by sharing market data through sophisticated data analytics. In collaboration with one of the world’s leading digital platforms, Alibaba’s Taobao Marketplace, we study how freely sharing advanced analytics by Taobao with its marketplace retailers affects their sales performance. Employing a synthetic difference-in-differences(SynthDiD) approach, we find: (1) adoption of advanced analytics resulted in a 30% increase in retailers’ sales revenue, as an addition to basic descriptive analytics; (2) market-level data analytics, which involves analyzing overall retailers’ data on the platform, generate more significant value than firm-level data analytics using only individual retailer data, have bigger impacts on smaller retailers than larger retailers, and demonstrate increasing return over time; (3) the mechanism analysis shows that the observed positive impacts are not attributed to changes in retailers’ pricing or advertising spending, but rather to product innovation; adopted retailers tend to release more new products and are more inclined to expand into new categories with less competition. Our findings contribute to the emerging global debate on data regulation and digital platforms, providing novel managerial and policy insights for digital platforms in efficiently utilizing non-rival customer data to empower small businesses.

Keywords: Data Sharing, Digital Platforms, Data Analytics, Data Regulation, SME Digital Transformation
1. Introduction

Data, considered the world’s most valuable resource\(^1\), is increasingly playing a critical role in the digital economy. It enables data-driven decision-making (Brynjolfsson and McAfee 2012) and enhances digital marketing capabilities in the marketplace (Wedel and Kannan 2016, Berman and Isareal 2022). Data possesses a nonrival characteristic, meaning it can be used concurrently by multiple entities without depletion, thus accelerating value creation (Jones and Tonetti 2020). Particularly in digital marketing contexts, customer behaviors are meticulously documented in data, enabling the derivation of substantial value. Employing a variety of analytical techniques, including rapidly iterated artificial intelligence models, diverse firms extract distinct values from these nonrivalrous gold mines.

However, despite the promising value of data sharing, practical implementation is hindered due to stringent data privacy and security regulations (Goldfarb and Tucker 2012, Varian 2018, Ke and Sudhir 2023). Moreover, most traditional firms, particularly small and medium enterprises (SMEs), have limited digital capability to collect, process, and use data, which becomes the biggest obstacle and most pressing issue for their digital transformation (OECD 2021).

On the other hand, digital platforms such as Amazon and Alibaba can collect unprecedented market-level data on all retailers’ and buyers’ behaviors in their ecosystems. Such market data can create huge economies of scope in data aggregation, i.e., the value of aggregated data from markets is higher than the sum of values of individual firm datasets (Martens 2021). Importantly, digital platforms have sophisticated data analytic capabilities to process and mine massive data to fuel their business models. Unlike traditional data aggregators, such as utility providers (e.g., telecom, bank) and marketing research firms (e.g., Nielsen, Kantar), digital platforms have detailed data directly related to retailers and are incentivized to share with them.

Data analytics are categorized into four levels: descriptive, diagnostic, predictive, and prescriptive (Wedel and Kannan, 2016). Descriptive analytics, such as statistical dashboards, are the most widely utilized and offer considerable value to firms. The more sophisticated tiers, collectively known as advanced

\(^1\) From the cover article of *The Economist*, issue May 6-12, 2017, “The world's most valuable resource is no longer oil, but data”.
analytics, safeguard sensitive information contained within the original data. Consequently, these levels provide substantial business value while upholding rigorous data privacy and security standards through data integration and advanced analytical techniques (Lismont et al. 2017, Anand and Lee 2023, Subramanian 2023).

Consequently, digital platforms are uniquely positioned to play a pivotal role in the digital economy by sharing market data with firms, particularly SMEs, through analytics. Through this approach, the value of data can be amplified and distributed to disadvantaged firms. It is urgent for managers and regulators to understand the value of sharing market data through data analytics for both SMEs and digital platforms.

Nascent studies find external data analytics positively impact firm performance (Berman and Isareal 2022), particularly SMEs (Bar-Gill et al. 2024). However, extant literature is silent on the role of digital platforms in market data sharing frameworks. We aim to evaluate such an approach through advanced data analytics. Specifically, we investigate the following research questions: (1) What impact does the sharing of data analytics by the digital platform have on retailer performance? (2) How do market-level data analytics, utilizing data on all sellers’ and buyers’ behaviors on the platform, differ from firm-level data analytics, using only a seller’s data, in their impacts? (3) What specific underlying mechanisms drive the impact of each type of data analytics?

We conduct a large-scale empirical study collaborating with the world’s leading digital platform, Alibaba’s Taobao online marketplace. Similar to Google Analytics and Amazon’s Brand Analytics, Taobao initiated an analytics program named Business Advisor in 2013. The program provides a free version dashboard with descriptive analytics from firm-level data for all retailers. It also offers a pro-version with advanced analytics from both firm- and market-level data charging a subscription fee for retailers. These advanced analytics leverage more sophisticated AI and machine learning models. The market-level data analytics use market data beyond the scope of any single retailer. Importantly, no type of analytic results can be used to identify any individual customer or firm to ensure no data privacy concerns. In May 2021, the platform changed the policy, making the advanced analytics pro-version free for all retailers. This provides us with an ideal quasi-experimental setting to examine the impacts of advanced analytics on the platform. We acquired a unique dataset including 100,000 retailers over a year from the Taobao platform.
and use a synthetic difference in difference (SynthDiD) approach to address the potential self-selection issue. From the estimation results, we show that, first, advanced analytics contribute to, on average, a 30% increase in sales for firms, on top of what is achieved through descriptive analytics. Second, market-level data analytics contribute more substantially to sales growth than firm-level data analytics, and show an increasing return in the long run, revealing a dynamic and sustained economic impact rather than a transient effect. Third, small retailers derive more significant benefits than large retailers, with the additional improvement primarily driven by market-level data analytics. Fourth, the growth firms experienced is mainly attributed to product innovation, highlighting the pivotal role of digital transformation.

This paper is the first empirical study to demonstrate the role of digital platforms in sharing market data through analytics and empowering small businesses. The findings provide important implications for scholars, practitioners, and regulators. The contribution of our findings is fourfold. First, we enrich the literature on the platform economy by providing new evidence on how digital platforms can leverage the nonrivalry property of data for broader economic benefits and to empower small businesses. Second, this research contributes to the ongoing debate on data sharing and privacy by exploring the role of digital platforms and data analytics. Third, we contribute to the literature on customer data analytics by studying advanced analytics from external market data. Fourth, our study uncovers that product innovation serves as a potential driver behind the sales growth of SMEs empowered by platform market data analytics, which suggests an important aspect of digital transformation for firms.

The remainder of the paper is organized as follows: Section 2 summarizes relevant literature. Section 3 details the empirical study design and data. Section 4 presents the overall main results and heterogeneous effects. Section 5 examines the underlying mechanism. Section 6 discusses implications for practice and regulation, then section 7 concludes the paper.

**2. Related Literature**

*The economics of data.* The emerging literature on the economics of data mainly studies the unique features of data and the role of data in markets and the economy (Marten 2021, Farboodi and Veldkamp 2023, Veldkamp and Chung 2024). For instance, Jones and Tonetti (2020) theoretically show broad usage or sharing of individual data results in significant social gains due to the nonrival nature of data. Bergemann
et al. (2022) analyze how the social value of individual data affects data acquisition and sharing in the market. However, discrimination or privacy violations are key issues (Goldfarb and Tucker 2012, Martin 2015). Echoing this concern, the European Union's General Data Protection Regulation (GDPR) was marked as a milestone for data regulation. Literature has largely found the privacy regulations from GDPR impair the economy (Johnson et al. 2023), including reduced investment in technology ventures (Jia et al. 2021), compromised advertising performance and revenue (Wang et al. 2024), increased consumer search efforts (Zhao et al. 2022), hindered the realization of potential data synergies (Gal and Aviv, 2020), and harmed smaller firms (Johnson et al. 2023). Addressing the conflict between data value and privacy concerns, Anand and Lee (2023) show that data analytics from machine learning algorithms can achieve the balance between accurate statistics and privacy protection. Our study contributes to the literature by proposing a market data sharing framework through platform data analytics programs as a potential solution for reconciling the economic benefit of data with the demand for privacy. We empirically demonstrate that such sharing, even under the current strict data regulation environment, improves firms' performance significantly, and such benefit is bigger for smaller businesses.

**Digital platforms.** Current literature on digital platforms mainly focuses on aspects such as two-sided markets and network externalities (i.e. Armstrong 2006, Tucker & Zhang 2010, Chu & Manchanda 2016, Brunswicker et al. 2019), and forecasts a “winner-take-all scenario” where the platform with the most users ultimately dominates the market (e.g., Besen and Farrell 1994, Caillaud and Julien 2003). In response to this concern, regulators from the European Union, and the U.S. started to scrutinize digital platforms from an antitrust perspective (Crémer et al. 2019, US House of Representatives, 2020). However, with the unprecedented amount of detailed market data on these platforms, digital platforms have the potential to empower a broad range of firms to unlock the social value of data (Galperti et al. 2023). Limited research has empirically examined the critical role of digital platforms in shaping markets and the economy. One exception is the study by Bar-Gill et al. (2024), which finds that the adoption of analytics dashboards from the eBay platform improves sellers’ data-driven decision-making and, consequently, revenue for SMEs on eBay. However, the analytics service examined in their study primarily focuses on descriptive analytics and largely relies on data at the individual firm level. In contrast, our study investigates how the policy of the
Taobao platform, which shares its advanced analytics service with sellers based on both firm- and market-level data on the platform, affects retailer performance.

**Customer data analytics.** With the prevalence of data in the marketplace, a large amount of marketing research has started to examine how customer data analytics affect marketing decisions and firm performance (e.g., Wedel and Kannan 2016, Bradlow et al. 2017). Simister et al. (2019) demonstrate the robustness of various widely used machine-learning methods in targeting prospective customers. A closely related study by Berman and Israeli (2022), finds a positive impact of adopting a retail descriptive analytics dashboard, based on firm-level data, on online retailer sales performance. Our study complements this research by examining the value of advanced data analytics based on both firm- and market-level data. Furthermore, we uncover different mechanisms underlying the impact of these two types of data analytics.

3. **Data and Empirical Strategy**

3.1 **Institutional Background**

We collaborated with the world’s leading online marketplace, the Alibaba Group’s Taobao e-commerce platform. Taobao initiated a data analytics program, *Business Advisor*, in 2013 with free and paid modules focused on different aspects of the business. Sample screenshots of the program dashboard are shown in Figure 1, and Table 1 briefly describes its major modules. To illustrate, a nutrition products seller on Taobao highlights the advanced analytics utility: the free dashboard provides descriptive statistics of their transactions and visitors, and the manager can have an overview of their firm data and identify fast-selling products to consider further promotions. For the firm analytics pro version, the manager obtains portraits of their customers, for instance, modern female office workers from first-tier cities, and recommendations of store page optimization based on the analysis of their visitors’ click trails; finally, the pro version's market analytics offers trending customer interests on the market like winter disease prevention. In response, they redesigned their product imagery to be more modern for younger demographics and developed new multi-nutritional products targeting the immune systems for a series of specific diseases.

Based on the specific modules, the pro version’s annual subscription fee varied from CNY 288 to CNY 19,800. The fees stayed stable and did not experience any gradual decrease or increase before the policy
change. We classify the free version as descriptive analytics and the pro version as advanced analytics, as those modules provide diagnostic, predictive, and prescriptive analytics with sophisticated algorithms. In addition, we distinguish the two types of advanced analytics by their data source, firm-level uses data from the individual seller, and market-level uses the customer and seller behavior data across the entire platform.

**Figure 1. Screenshots of Business Advisor Dashboard**

![Figure 1. Screenshots of Business Advisor Dashboard](image)

<table>
<thead>
<tr>
<th>Name</th>
<th>Firm-level Data Analytics</th>
<th>Market-level Data Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brief Description</strong></td>
<td><strong>Product Compass</strong></td>
<td><strong>Traffic Cross</strong></td>
</tr>
</tbody>
</table>
| **Firm-level Data Analytics** | Analyze a retailer’s listed products.  
*Free version (Descriptive Analytics)*: preliminary statistics and visualizations for product performance, such as views, carts, and sales.  
*Pro version (Advanced Analytics)*: suggested store and product optimization, new product evaluation, problem detection, etc. | Analyze the retailer’s customers.  
*Free version (Descriptive Analytics)*: preliminary statistics on visitors and purchase behaviors.  
*Pro version (Advanced Analytics)*: customer portraits, high-potential visitor detection, traffic acquisition strategies, etc. | *Pro Version Only (Advanced Analytics)*: This module analyzes the comprehensive market data. The analytical results are highly aggregated, such as trending search terms, brand popularity rankings, categories searching index, crowd portraits, etc. |

### 3.2. The Quasi-Experimental Design

There are several identification challenges in measuring the impact of advanced analytics on retailers. First, as access to advanced analytics incurs subscription fees to the pro version, more productive firms may have more significant funding to invest in data analytics, potentially confounding the causal interpretation. Second, firms usually consider initiating advanced techniques when anticipating a high return, leading to endogenous adoption timing. Third, there might be other unobserved factors that affect the decision to adopt, thus incurring selection bias.
In May 2021, Taobao implemented a policy change that waived fees for all Taobao retailers on all professional modules. The Business Advisor did not undergo any significant redesign or functional update during our observation periods, and no other alternative analytical programs were available to the retailers. Therefore, the sole distinction between pre and post-policy change was eliminating the subscription cost to retailers. Meanwhile, the information on free access to advanced analytics was shown in prominent locations on the web pages to ensure the retailers were informed weeks before the launch date. This policy change provided an ideal quasi-experimental setting for our empirical study and enabled us to identify several study cohorts to address the above challenges. We construct the treatment group of retailers that have descriptive analytics experience and adopted advanced analytics during the period of policy change, therefore, their adoption decisions are restrained by cost and directly linked to the policy change. Meanwhile, their adoption timing is attributed to the exogenous events of cost removal and the accompanying publicity campaign. Furthermore, we use a synthetic method and several robustness checks to manage the potential self-selection bias.

3.3. Empirical Strategy

After the platform launched the free sharing campaign for its Business Advisor data analytics program, retailers adopted the program in a staggered fashion, which makes it complicated to estimate the treatment effects (Lin et al. 2020, Chaisemartin and d’Hault-foueille 2020, Adjerid et al. 2023). Besides, retailers may employ various levels and types of analytics concurrently, making the inference for contribution difficult. By utilizing study cohorts for our analysis, we are able to obtain unconfounded results for the research question we aim to examine. We establish each study cohort based on the retailers' engagement with data analytics services during specific periods of the year. Consequently, this ensures that the treatment and control groups are comparable in terms of their experience with analytics at the time of the intervention. Figure 2 depicts the quasi-experimental design. Cohort G1 are retailers that have not adopted any analytics in our observation periods; Cohort G2 are retailers that first adopted descriptive analytics in May 2021 and stayed at this level in our observation periods; By analyzing these two cohorts, we can obtain the impact of adopting descriptive analytics for retailers without any analytics experience, this result serves as a baseline of our study. Cohort G3 consists of retailers who adopted descriptive analytics prior to our observation.
periods and maintained this level throughout the entire period; Cohort G4 includes retailers who similarly adopted descriptive analytics before our observation periods but upgraded to advanced analytics during the policy change period (May 2021). Consequently, retailers in both cohorts G3 and G4 possess at least one quarter's experience with descriptive analytics, allowing us to isolate the impact of advanced analytics over and above descriptive analytics.

We define Cohort G4 to exclude retailers that adopted the program between June and December; this method ensures clean inference by removing the potential confounder of endogenous adoption timing. Notably, Cohort G4 comprises 61.27% of all qualified advanced analytics adopters. We attribute the predominant drivers of adoption in May primarily to the policy change that significantly reduced access costs to zero. Table 2(b) presents the descriptive statistics across cohorts.

**Figure 2. The Quasi-Experimental Design**

Further, with our main analysis between study cohort G3 and G4, identifying the impacts of different types of advanced analytics is challenging; cohort G4 retailers did not uniformly adopt firm-level or market-level analytics as they upgraded; instead, they did so in a staggered manner. Many firms adopted both types of analytics at the beginning; for firms that initially adopted firm-level, some kept the status quo, and some gradually added market-level analytics in the following periods. Firms that started with market-level...
analytics alone are less than 1% of G4 and are negligible. Previous literature generally aggregates the separate estimated treatment effects (Goodman-Bacon 2018, Callaway and San-t’Anna 2021, Wooldridge 2021). However, we aim to show comparable results for advanced analytics users with additional market-level analytics and those without to evaluate their contribution to the impact of advanced analytics. Therefore, we further constructed two sub-cohorts from cohort G4. Sub-cohort G4f retailers only adopted firm-level advanced analytics until the end of 2021, and those in sub-cohort G4m adopted additional market analytics since the first month of the upgrade. These two sub-cohorts of retailers are not entangled in post-treatment periods, thus ensuring a clean comparison of contribution to the impact of advanced analytics.

3.3. Data

We acquired a unique dataset of 100,000 randomly selected active retailers from Taobao. We first exclude the retailers that had paid and used the pro-version before May 2021, as they encountered confounding factors, which yielded 87,494 retailers. The records are in monthly granularity, specifically, our panel data includes three parts.

**Retailer characteristics.** The data encompasses details like each retailer's location, age, consumer rating, platform-assigned rating, and industry sector, including categories such as clothing and consumer electronics.

**Retailer performance and decisions.** Monthly sales revenue data for each retailer were gathered, representing the primary metric for evaluating firm performance. We also obtained data that represent each retailer’s market performance each month, including the number of page views, unique visitors, number and sales of new and repeat customers, sales from the bestselling product, number of paid customers, number of items sold and added to the shopping cart, average transaction value, advertising spending, SKU counts, and list of categories of the products.

**Analytics usage.** Business Advisor is structured with distinct modules requiring separate subscriptions rather than a single consolidated dashboard. The data records the total login times for each retailer each month for the specific analytics module. We aggregate monthly login times of the free versions of Product Compass and Traffic Cross as descriptive analytics usage and corresponding pro-versions are considered.
firm-level advanced analytics usage. We use the login times of Market Insights as market-level advanced analytics usage. The usage of both firm-level and market-level analytics is aggregated to represent the total advanced analytics usage. Table 2 presents the summary statistics for the overall sample and by study cohorts.

Table 2. Summary Statistics

(a) All samples

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Std.</th>
<th>Median</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>41,524.96</td>
<td>110,096.83</td>
<td>9,794.45</td>
<td>87,494</td>
</tr>
<tr>
<td>Avg.Price</td>
<td>168.64</td>
<td>567.60</td>
<td>50.13</td>
<td>87,494</td>
</tr>
<tr>
<td>Daily Page Views</td>
<td>1,061.51</td>
<td>5,003.12</td>
<td>154.07</td>
<td>87,494</td>
</tr>
<tr>
<td>Daily Unique Visitors</td>
<td>388.88</td>
<td>1,610.29</td>
<td>60.26</td>
<td>87,494</td>
</tr>
<tr>
<td>ATV</td>
<td>514.12</td>
<td>2,667.55</td>
<td>146.52</td>
<td>87,494</td>
</tr>
<tr>
<td>Buyers Count</td>
<td>283.91</td>
<td>1,636.66</td>
<td>55.33</td>
<td>87,494</td>
</tr>
<tr>
<td>Items Sold</td>
<td>17,232.80</td>
<td>1,468,784.28</td>
<td>192.50</td>
<td>87,494</td>
</tr>
<tr>
<td>SKU Count</td>
<td>384.99</td>
<td>3,081.74</td>
<td>35.58</td>
<td>87,494</td>
</tr>
<tr>
<td>Repeat Customers Sales</td>
<td>13,876.90</td>
<td>53,549.63</td>
<td>1,969.67</td>
<td>87,494</td>
</tr>
<tr>
<td>New Customers Sales</td>
<td>25,277.08</td>
<td>75,955.82</td>
<td>4,925.92</td>
<td>87,494</td>
</tr>
</tbody>
</table>

Notes: ATV stands for Average Transaction Value.

(b) By study cohorts

<table>
<thead>
<tr>
<th>Variables</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>13,682.52</td>
<td>35,626.57</td>
<td>48,084.88</td>
<td>107,531.15</td>
</tr>
<tr>
<td>(45,743.12)</td>
<td>(89,380.82)</td>
<td>(122,679.05)</td>
<td>(193,705.99)</td>
<td></td>
</tr>
<tr>
<td>Avg.Price</td>
<td>145.01</td>
<td>140.35</td>
<td>177.33</td>
<td>207.44</td>
</tr>
<tr>
<td>(578.99)</td>
<td>(358.90)</td>
<td>(508.85)</td>
<td>(722.83)</td>
<td></td>
</tr>
<tr>
<td>Daily Page Views</td>
<td>288.82</td>
<td>774.48</td>
<td>1,442.66</td>
<td>2,861.40</td>
</tr>
<tr>
<td>(1,407.76)</td>
<td>(2,951.95)</td>
<td>(6,911.49)</td>
<td>(8,291.28)</td>
<td></td>
</tr>
<tr>
<td>Daily Unique Visitors</td>
<td>130.61</td>
<td>294.50</td>
<td>494.27</td>
<td>1,036.97</td>
</tr>
<tr>
<td>(726.21)</td>
<td>(1,016.69)</td>
<td>(1,922.00)</td>
<td>(2,886.51)</td>
<td></td>
</tr>
<tr>
<td>ATV</td>
<td>530.78</td>
<td>476.41</td>
<td>529.44</td>
<td>474.12</td>
</tr>
<tr>
<td>(2,809.26)</td>
<td>(1,712.45)</td>
<td>(4,386.63)</td>
<td>(1,740.97)</td>
<td></td>
</tr>
<tr>
<td>Buyers Count</td>
<td>141.47</td>
<td>229.33</td>
<td>269.24</td>
<td>749.33</td>
</tr>
<tr>
<td>(1,489.30)</td>
<td>(857.97)</td>
<td>(1,194.87)</td>
<td>(2,345.00)</td>
<td></td>
</tr>
<tr>
<td>Items Sold</td>
<td>24,923.89</td>
<td>3,071.40</td>
<td>39,441.89</td>
<td>3,621.25</td>
</tr>
<tr>
<td>(2,085,932.10)</td>
<td>(26,574.35)</td>
<td>(2,022,341.15)</td>
<td>(28,773.03)</td>
<td></td>
</tr>
<tr>
<td>SKU Count</td>
<td>605.64</td>
<td>365.11</td>
<td>249.91</td>
<td>222.69</td>
</tr>
<tr>
<td>(3,918.22)</td>
<td>(3,206.08)</td>
<td>(1,516.86)</td>
<td>(3,333.14)</td>
<td></td>
</tr>
<tr>
<td>Repeat Customers Sales</td>
<td>6,476.26</td>
<td>13,970.73</td>
<td>17,989.53</td>
<td>27,954.38</td>
</tr>
<tr>
<td>(29,354.97)</td>
<td>(48,058.67)</td>
<td>(66,420.20)</td>
<td>(75,784.42)</td>
<td></td>
</tr>
<tr>
<td>New Customers Sales</td>
<td>7,205.68</td>
<td>21,655.15</td>
<td>30,094.67</td>
<td>79,575.94</td>
</tr>
<tr>
<td>(24,985.73)</td>
<td>(59,239.58)</td>
<td>(77,967.51)</td>
<td>(148,895.71)</td>
<td></td>
</tr>
<tr>
<td>Number of Retailers</td>
<td>34,493</td>
<td>3,368</td>
<td>15,714</td>
<td>5,566</td>
</tr>
</tbody>
</table>

Notes: ATV stands for Average Transaction Value. The values are means, and standard deviations are in parenthesis.

3.4. Econometric Model

To obtain robust estimations and address selection bias, we use the SythDiD approach, which is suitable for unified treatment time and large panels of data spanning various industries (Arkhangelsky et al.
Combining the advantages of synthetic control and DiD, SynthDiD yields robust estimations without a strong reliance on pre-treatment parallel trends. SynthDiD is more generalized and conservative for our large dataset of firms across many industries. Retailers in our randomly selected sample have widely differentiated business models; for instance, some stores sell zipper pullers for about CNY 1 each, and some stores sell luxury watches for over CNY 100,000 unit price. Therefore, as a unified measurement, we consider sales revenue as the most informative variable and build the synthetic control group. The synthetic control method optimizes weights for the control group to enable the weighted average of their outcomes to predict the outcome if the treatment group had not received it (Abadie 2021). However, this method is typically adopted when dealing with a small number of study subjects or constructing a synthetic match for each unit in the treatment group (Acemoglu et al. 2016). SynthDiD optimizes for both control units and pre-treatment periods to minimize the mean squared errors of the estimated average treatment effect.

For each pair of control and treatment groups, we first apply the optimization method Arkhangelsky et al. (2021) proposed to estimate synthetic weights for each unit in the control group during pre-treatment periods. We then employ model (1) to obtain the treatment effect.

\[
(\tau, \mu, \alpha, \beta) = \arg \min_{\tau, \mu, \alpha, \beta} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \omega_i \lambda_t \right\},
\]

(1)

Where \(Y_{it}\) is the sales revenue for retailer \(i\) at month \(t\), and \(W_{it}\) indicates whether the retailer at month \(t\) has adopted the corresponding program. \(\omega\) and \(\lambda\) represent optimized weights for each control unit and pre-treatment period. Finally, \(\tau\) is the target average treatment effect on treated (ATT), indicating the impact of analytics adoption on the treatment group. We use the jackknife method (Miller 1974) to obtain the standard error for \(\tau\).

4. Results

4.1. Impacts of Data Analytics Adoption

We carefully selected pairs of study cohorts as control and treatment groups to tease out the impacts of different types of data analytics. This section details the estimation results for overall and heterogeneous effects using the SynthDiD method.

**Descriptive analytics.** We use cohort G1 as the control group and G2 as the treatment group to assess...
the effects of descriptive analytics. Column (1) in Table 3 reports the estimation results. The ATT for the impact of descriptive analytics is 35.66% \((=\exp(0.305)-1)\), in 8 months after adoption, equating to an economic impact equivalent to CNY 2,826.8 per month for a median retailer in the sample.

**Advanced analytics.** We use cohort G3 as the control group and cohort G4 as the treatment group to study the impacts of advanced analytics. We estimate the model presented in Equation (1), and use the new weights to calculate \(\tau\) and standard deviation with the jackknife method. Column (2) in Table 3 presents that the ATT for the overall advanced analytics impact is 30% \((=\exp(0.263)-1)\) for the median retailer in the sample, equivalent to CNY 11,359.8 per month\(^2\). Figure 2 shows the dynamic impact of advanced analytics in each period, indicating an undiminishing return in the long run. Web Appendix A details the estimation procedure and explicit results for each period.

**Firm- vs. Market-level analytics.** We aim to tease out the contribution of firm-level and market-level analytics from the overall advanced analytics. Berman and Israeli (2022) creatively compare three methods of computing subgroup effects using SynthDiD, and we adopt a similar approach. We fix weights \(\omega\) and \(\lambda\) from the last estimation and compute the treatment effect \(\tau\) separately with sub-cohorts G4f and G4m. This allows for a clean comparison of effects as both sub-cohorts use the same control group, cohort G3. From the dynamic analysis, we observe that the difference in pre-treatment sales is small. Web Appendix A details the estimation process for each period.

As shown in bottom half of Table 3, the estimation result shows a 19.8% \((=\exp(0.178)-1)\) increase in sales revenue for those relying solely on firm-level analytics (G4f), and a 34.45% \((=\exp(0.296)-1)\) increase in sales revenue for the group using both market-level and firm-level analytics (G4m). The disparity in these two impacts is attributed to market-level analytics. Figure 3(a) presents the dynamic estimation of the two sub-cohorts of retailers. Notably, the returns from firm-level analytics are decreasing in the long run; in contrast, the returns for the group using both market- and firm-level analytics remain stable or even increasing in the long run, suggesting that market-level data analytics exhibits an increasing return. This behavior of market analytics aligns with similar properties observed in capital and R&D input (Farboodi and Veldkemp, 2022). Given that the data is collected from the market by the platform and shared across

---

\(^2\) The median average monthly sales revenue for cohort 2 and 4 retailers in pre-treatment periods are CNY 7926.6 and CNY 37761.9.
firms, the increasing returns can also be attributed to the non-rivalry of data (Jones and Tonetti, 2020; Veldkamp and Chung, 2024). We further examine the unique mechanisms of market analytics in section 5.

<table>
<thead>
<tr>
<th>Table 3. Impact of Different Types of Data Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive Analytics</td>
</tr>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>log(sales)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Analytics Types</td>
</tr>
<tr>
<td>Firm-Level</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Market-Level</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note: Significance level: p<0.1 (*), p<0.05 (**), p<0.01 (***). Robust standard errors appear in parentheses.

Figure 2. Overall Average Treatment Effects at Each Period.

4.2. Heterogeneous Effect across Retailer Size

Retailers on digital platforms span various industries; thus, we must find a coherent measure representative of retailer size. Following Han et al. (2022), which studied retailers on the same platform as ours, we use the mean pre-treatment sales and define the top 25% as large retailers and the bottom 75% as small retailers. In our dataset, the average annual revenue is CNY 457,998 for small retailers and CNY 3,786,305 for large retailers. The small retailers in our sample align with the SMEs in Bar-Gill et. al (2024),

3 The average exchange rate is 1 CNY to 0.155 USD in the year of 2021. The annual revenue is equivalent to USD 70,989 for small retailers and USD 586,877 for large retailers.
as the vast majority of the SMEs in their sample have annual sales on eBay between $10,000 and $100,000.

Columns (1) and (2) in Table 4 present the estimation results. Small retailers received a higher impact (32.4%=exp (0.281)-1) than large firms (23.2%=exp (0.209)-1) from advanced analytics. The improvement magnitude for small firms is greater, indicating a significant empowerment for those lacking both capacity and data. Regarding dynamic effects, as shown in Figure 3(b), small firms show a steeper increasing return curve from data analytics over time, showing a rapid progression as their information gap filled. Still, the returns from data analytics for large firms become more prominent over time, but in a flatter learning curve. This pattern aligns with the findings of Tambe and Hitt (2012), who argue that IT returns materialize more slowly in large firms. However, for smaller firms, the short-run contribution of IT to output is similar to the long-run output contribution.

Table 4. Heterogeneous Impact of Advanced Analytics

<table>
<thead>
<tr>
<th>Firm Size</th>
<th>Small (Bottom 75%)</th>
<th>Large (Top 25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Sales)</td>
<td>0.281*** (0.019)</td>
<td>0.209*** (0.017)</td>
</tr>
<tr>
<td>Treated Units</td>
<td>4,174</td>
<td>1,391</td>
</tr>
</tbody>
</table>

Analytic Types Cross Effect

<table>
<thead>
<tr>
<th>Analytic Type</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm-Level</td>
<td>0.246*** (0.025)</td>
<td>0.279*** (0.031)</td>
</tr>
<tr>
<td>Market-Level</td>
<td>0.442*** (0.044)</td>
<td>0.303*** (0.046)</td>
</tr>
</tbody>
</table>

Note: Significance level: *p<0.1, **p<0.05, ***p<0.01

Robust standard errors appear in parentheses.

We further divide small vs. large retailers into more sub-groups with their adoption of market analytics to compare the cross effects with analytic types. The bottom half of Table 4 presents the estimated results. An important finding is that market-level data analytics contributes to the higher overall growth of small retailers. When adopting firm-level analytics only, small retailers have less improvement compared to large firms (27.89%= exp (0.246)-1 vs. 32.18%= exp (0.279)-1); however, when additionally adopting market-level analytics, small retailers have a much higher improvement (55.58%= exp (0.442)-1 vs. 35.39%= exp (0.303)-1). Small retailers usually generate less data from their operations to be analyzed, resulting in a less
effective data feedback loop compared to large retailers, and often fall into data poverty traps (Farboodi and Veldkamp, 2022); with external market data, this information gap is bridged. This finding provides important evidence for the market data sharing feature of digital platforms to empower small businesses.

**Figure 3. Average Treatment Effects on Analytics Type and Retailer Size in Each Period.**

![Graphs showing average treatment effects by analytics type and retailer size](image)

(a) By Analytics Type               (b) By Retailer Size

**4.3. Robustness Checks**

To address the potential self-selection bias and other confounding effects of unobservables, we analyze the robustness of our main finding through several alternative methods.

**Alternative cohorts.** We consider retailers adopting advanced analytics in May 2021 were willing but previously constrained by cost, with their decisions catalyzed by the free access policy change. Since the treatment was not randomly assigned, other unobserved factors may have influenced the timing of adoption. To mitigate these concerns, we replicate the analysis for retailers whose first adoption occurred in subsequent periods following the policy change. We establish four additional study cohorts representing retailers that have experience with descriptive analytics and first adopted advanced analytics in June, July, August, and September 2021, along with corresponding control groups for each. We use the same methods to estimate both the overall and heterogeneous impacts for each alternative cohort. The Web Appendix B
detailed procedure and results, suggest that the effects of advanced analytics, as well as the impact of different types of analytics across these alternative cohorts, are largely similar.

**Propensity Score Matching.** As a robustness check for self-selection bias, we utilize the propensity score matching (PSM) method to match retailers in study cohorts G4 and G3 in a robust manner. We use stores’ demographics, performance, and descriptive analytics usages in pre-treatment periods to implement the matching, see Web Appendix C for detailed procedures, validation tests, and estimation results. The parallel trend assumption is satisfied, and after a balanced matching, our match sample includes 4,640 firms in both groups. Column (1) of Table 5 presents a significant 33% (exp(0.285)-1) positive impact on sales revenue, equivalent to an economic benefit of CNY 9742.7 for treated firms. The result closely aligns with our primary analysis.

**Instrumental Variable Analysis.** To further test the robustness of our main finding, we use an instrumental variable analysis with the change in access cost of advanced analytics as the instrument. This instrument is a binary variable and is 0 for all retailers before the policy change and 1 thereafter. The free access indicator affects the retailers’ adoption decision but is uncorrelated with their performance as it is an exogenous event. We include all retailers in our sample besides those who never adopt descriptive analytics to ensure the estimated impact is indeed for advanced analytics as an addition to descriptive analytics. We utilize the Heckman two-step approach with a staggered difference-in-difference as the second stage model, see Web Appendix D for detailed procedures and results. Column (2) in Table 5 shows the IV analysis results with a significant revenue increase of 22.9% (exp(0.206)-1), which aligns with our main analysis.

**Pioneers as Controls.** A potential concern of the control group in our main analysis is unobservable factors to keep them from adopting advanced analytics even if the access cost is 0. In this analysis, we exclude those nonadopters and instead use retailers that paid the subscription fees to adopt advanced analytics, which we call pioneers, as the control group. Retailers in the alternative control group have access to advanced analytics in the entire period of our sample, therefore, the treatment in this analysis is referred to
as filling the gap of data sharing through advanced analytics. We estimate synthetic weights for the new control group and estimate the SynthDiD model, the result is shown in column (3) of Table 5. Filling the gap of data sharing has a significant impact of 16.3% \((\exp(0.151)-1)\) increase in sales revenue. Note that the pioneers are larger stores, thus the magnitude is smaller in our main analysis.

In addition, we trimmed cohort G4 by removing the outliers defined by their mean sales outside the 1 to 99 percentile. Then we re-do the synthetic weights estimating for G3 and then the SynthDiD model, as shown in the last column of Table 5, the impact is similar to the main analysis.

**Table 5. Estimates on Adoption Impacts**

<table>
<thead>
<tr>
<th></th>
<th>Difference in Difference with Propensity Score Matching (1)</th>
<th>IV Model with Two-Step Heckman (2)</th>
<th>Alternative Control Group (3)</th>
<th>Trimmed Treatment Group (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\log(sales))</td>
<td>0.285***</td>
<td>0.206***</td>
<td>0.151***</td>
<td>0.275***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.012)</td>
<td>(0.016)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Num. Retailers</td>
<td>9,280</td>
<td>57,669</td>
<td>11,766</td>
<td>20,706</td>
</tr>
</tbody>
</table>

*Note: Significance level: \(p<0.1\) (*), \(p<0.05\) (**), \(p<0.01\) (***)

Robust standard errors appear in parentheses.

5. Mechanisms

We investigate three possible mechanisms that may contribute to the increase in sales from firm-level and market-level analytics independently. First, adjusting prices or conducting price promotions seems a straightforward strategy once retailers have gained a deeper understanding of their customers and competition status. Second, retailers might increase advertising to attract new customers and boost sales. Third, retailers might introduce new products tailored to their customers after gaining insights from firm and market analytics.

5.1. How Data Analytics Drive Firm Decisions

We use different measurements representing the three essential marketing decision variables as
dependent variables and estimate them using SynthDiD. Table 6 presents the results.

**Table 6. Impact on Decision Variables by Firm and Market Analytics.**

<table>
<thead>
<tr>
<th>Overall</th>
<th>(1) Log(Average Price) Firm</th>
<th>(2) Ad. Spending Firm</th>
<th>(3) Log(SKU count) Firm</th>
<th>(4) Category Expansion Firm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log(Average Price) Market</td>
<td>Ad. Spending Market</td>
<td>Log(SKU count) Market</td>
<td>Category Expansion Market</td>
</tr>
<tr>
<td>ATT</td>
<td>0.001</td>
<td>0.012</td>
<td>0.083***</td>
<td>0.192*</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.013)</td>
<td>(0.008)</td>
<td>(0.101)</td>
</tr>
<tr>
<td>Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>0.005</td>
<td>0.019</td>
<td>0.068***</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.016)</td>
<td>(0.010)</td>
<td>(0.114)</td>
</tr>
<tr>
<td>Large</td>
<td>−0.014</td>
<td>−0.010</td>
<td>0.134***</td>
<td>0.480***</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.022)</td>
<td>(0.011)</td>
<td>(0.143)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the price decision, we use the average price as the measurement. As Column (1) in Table 6 shows, we find no significant price change resulting from either firm or market analytics. This suggests that the increase in sales revenue is not attributable to retailers’ price changes. For the advertising decision, we obtained data on retailers’ advertising costs on the platform from the platform. Due to the business secret concern, these costs were normalized and categorized into six levels from 0 to 5. We measure the advertising change in levels from the previous period for each period. For instance, if a retailer has a spending level 3 in the current period and level 2 in the last period, it will receive a positive shift of +1. As Column (2) in Table 6 presents, the insignificant estimation results indicate that there are no significant impacts from each type of advanced analytics on advertising spending levels. This suggests that retailers’ sales increase might not be caused by changing advertising spending.

Finally, retailers might leverage analytics insights for new product innovation. We examine firms’ new product innovation from two aspects. First, we use SKU counts to measure the extent of releasing new products. Second, we inspect whether the new products target unmet demands uncovered from advanced analytics by measuring product category expansion. Our dataset records a list of categories corresponding to the retailer’s products, such as menswear, stereo speakers, toiletries, etc. For each period, we count the number of new categories that were first added; each newly added category is weighted to reflect its market competition. The weight is higher if the category is served by fewer firms and vice versa; we define it as:

\[
(1 - \frac{N_{kt}}{N_{lt:(k\neq l)}})^2, \quad (2)
\]
Where $N_{kt}$ is the number of retailers that contain category $k$ at period $t$, and $N_{it:k\in I}$ is the total number of retailers in the industry $I$ to which category $k$ belongs. The weighted indicator measures the extent and effectiveness of category expansion that fills unmet customer demands.

As shown in columns (3) and (4) of Table 6, we find positive impacts of both firm and market analytics on both SKU count and category expansion. More importantly, market-level analytics have stronger effects on both types of new product innovation. Furthermore, the value from market-level analytics is more important for small than for large retailers. Small retailers do not significantly expand into new categories with only firm-level analytics, and with market-level analytics, they significantly boost their efforts in such expansion and release more SKUs, suggesting the external market data stimulates greater product innovation to address unmet market demand for small retailers than for large retailers, leading to higher sales revenue. In contrast, large retailers choose different strategies with and without market analytics. With only firm analytics, they tend to release new products in new categories. With market analytics, they turn to within-category product expansion with a higher increase in SKU count. To verify this pattern, we find, on average, the SKU count per category from large retailers in the firm analytics group is reduced by 0.486 after adoption but increased by 3.338 after adoption with market analytics. This suggests that market analytics play different roles in product innovation for small and large retailers. One possible explanation is that large retailers obtain insights into their large customer base from firm analytics, thus want to extend to new categories to meet their demands in other areas; meanwhile, they are usually category leaders, and external market information provides a clearer view of market competition, prompting them to focus on product innovation within their advantageous categories.

### 5.2. How Data Analytics-driven Product Innovation Affects Customer Behavior

How can retailers’s data analytics-driven product innovation decisions affect customer behaviors, further resulting in retailer sales changes? We examine various mediating variables that reflect such changes. As shown in Table 7, first, market analytics generate higher growth than firm analytics in all customer behavior variables for all retailers, suggesting it is more instrumental in helping retailers discover potential demand and increase sales from both existing and new customers in the market. Second, such impact of market analytics is higher for small retailers than for large ones. Third, for small retailers, market analytics
have positive impacts on both repeat and new customer counts and sales, while firm analytics does not have a significant impact on repeat customer count. This suggests that with the customer insights market analytics provide, small retailers can attract both new and existing customers by expanding new products into new categories. Finally, while large retailers’ cross-category new product expansion strategy from firm analytics only attracts new customers, their within-category new product expansion strategy from market analytics increases both existing customers’ repeat purchases and new customers’ numbers and purchases.

6. Business and Policy Implications

Our findings offer key managerial and policy implications. First, for digital platforms, our findings demonstrate the economic benefits derived from integrating advanced analytics with market data for retailers. Digital platform managers will find it helpful as they assess their services, considering that many currently perceive analytics primarily as value-added services. Nevertheless, in contrast to relying solely on subscription fees, managers of digital platforms are advised to harness their market data to empower a wider range of firms. By doing so, the intrinsic value of data can be multiplied within the ecosystem, thereby enhancing the platforms' competitive edge through more vibrant trading activities and the elevation of seller performance.

Secondly, for firms within the retail sector, discerning customer needs is a pivotal growth determinant.
(Foster et al., 2016). Our research reveals that the utilization of advanced analytics in conjunction with market data enables firms to pinpoint unmet customer demands, thereby fostering more efficient and effective product innovation. Our study advocates for firms, particularly smaller ones, to adopt market data analytics in new product development, as external data can bridge the information gap with larger competitors. Firms should assess the potential benefits against the costs of accessing such data and consider the expansion of digital talents to amplify the value.

Finally, for regulators, our study joins the ongoing debate concerning personal data and the regulation of digital platforms. We explore the emerging role of digital platforms, which, with their prowess in collecting market data and their analytical capabilities, can serve as central hubs for the aggregation, amplification, and equitable distribution of data value. Such data sharing practices do not compromise individual privacy and can benefit disadvantaged firms. Digital platforms can function as gatekeepers within the data governance landscape, responsibly protecting identifiable information while facilitating the widespread dissemination of its value among firms. Policymakers will find our analysis beneficial in assessing regulatory frameworks.

7. Conclusion

We conducted a large scale empirical study on market data sharing through data analytics and found a 30% increase in sales revenue for retailers adopting the advanced analytics provided by the digital platform. Market-level analytics contribute more value, and the shared external market data is the key to empowering small businesses. Importantly, the growth stems from product innovation rather than changes in pricing or increased advertising spending. Our results provide concrete evidence for data sharing and governance dilemmas, which are rather important in the digital economy.

First, we provide a potential solution to amplify the value of market data with privacy concerns addressed; thus, digital platforms can utilize their advantage to multiply the value of data due to the non-rivalry property. While many digital platforms still regard analytics as value-added services for subscription fees, empowering firms on the platform is much more valuable in shaping the platform’s competitive advantage. Second, we show that shared market data allows retailers to discover and target customer demands more precisely, leading to effective innovation. Therefore, sharing market data with a broad range
of retailers can be a sustainable growth model for the ecosystem and the market in general, as innovation can lead to an increase in total customer demands. Third, our study presents the novel role of digital platforms in empowering small businesses. Platforms share market data and analytics capacity neutrally, thus benefiting disadvantaged firms. As the information gap is bridged by the platform, small retailers show much higher growth in magnitude than large retailers. Platforms should consider using it for a more vibrant ecosystem, and regulators can consider it as a method to create a more balanced market.

In this growth framework, digital platforms are indispensable, being the sole entities capable of and willing to share data and capacity with firms. Regulators may find our study useful in deliberating the role of digital platforms in the data economy, as they hold unprecedented customer data and can create multiplied values under regulation. Digital platforms act as gatekeepers in the data governance landscape, responsibly safeguarding identifiable data while disseminating its value across firms.

To our knowledge, this paper represents the first empirical study that systematically explores the impact of market data sharing through analytics on digital platforms. Our findings are essential for developing the data economy and indicating the importance of data sharing to disadvantaged firms. However, our results are based on one-year observation and impacts in longer terms are demanded for future research.

Reference


